0

N00014-93-1-0285

372 Three Seven Two

404558

0

QUANTUM COHERENCE AND REALITY

Dr. Yakir Aharonov

QUANTUM COHERENCE
AND
REALITY

In Celebration of the 60th Birthday of Yakir Aharonov

International Conference on
Fundamental Aspects of Quantum Theory

University of South Carolina, Columbia

10 – 12 December 1992

Editors

Jeeva S Anandan & John L Safko

University of South Carolina

World Scientific
Singapore • New Jrrsey • London • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.
P O Box 128, Farrer Road, Singapore 9128
USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

QUANTUM COHERENCE AND REALITY
In Celebration of the 60th Birthday of Yakir Aharonov

ISBN 981-02-2117-7

Printed in Singapore.

PREFACE

On December 9-12, 1992 over 150 scientists from around the world gathered in Columbia to celebrate the sixtieth birthday of Yakir Aharonov. The major portion of this celebration was a three day conference on the Fundamental Aspects of Quantum Theory. This volume is the proceedings of that conference and a brief biographical sketch of Yakir Aharonov as presented by Alex Pines after the banquet.

Among the topics discussed were the Aharonov-Bohm effect, geometric phases, gauge fields, black holes, quantum gravity, non-locality and geometry, spin and statistics, phenomenology, and quantum reality. These topics were chosen since they are all areas in which Yakir Aharonov has made contributions and suggestions.

Years ago developments in the fundamentals of quantum theory were primarily of interest only to theoreticians. Topics such as quantum gravity, non-locality and geometry, and black holes are still with us today; however, as can be seen from the table of contents, applications abound. Experiments have been performed showing flux lines, quantum interferometers are in use, and condensed matter applications and statistical applications exist. Recent satellite data provides information on black holes. The Aharonov-Bohm effect is now a laboratory phenomenon. Yakir Aharonov has recently demonstrated the reality of the wavefunction for a single particle.

In the years since the Aharonov-Bohm effect was proposed, Yakir Aharonov has made important suggestions and contributions to many areas related to fundamental interpretation of quantum theory. He has always taken the viewpoint that quantum theory must be studied to develop the necessary intuition to be able to understand what the theory is really telling us. Without this intuition we will often not ask the "right" question, and hence, misinterpret the basic nature of reality. That is, if we ask classical questions, we will see only some aspects of quantum theory. Intuition will enable us to ask the proper quantum question to discover the full implication of the theory. We dedicate this volume to him.

We had planned to have David Bohm, FRS, as a speaker at these sessions and to help honor his former student. We deeply regret his untimely death. He was a great physicist with a deep understanding of quantum theory and a humanistic person with a wide range of interests.

We express our appreciation to the aid provided by the members of the Scientific Advisory Committee: Michael Berry (Bristol), David Bohm (London), Roger Penrose (Oxford), Norman Ramsey (Harvard), Charles Townes (Berkeley), John Wheeler (Princeton) and Chen Ning Yang (Stony Brook). We also express out sincere appreciation to the other members of the Local Organizing Committee: Chi-Kwan Au, Frank Avignone, Richard Creswick, Horacio Farach, James Knight, Pawel Mazur, and Carl Rosenfeld. Without the help of both of these groups, this conference would not have been possible. We also gratefully acknowledge the generous support for this conference provided by President Palms of the University of South Carolina, the National Science Foundation, the Department of Energy, the Office of Naval Research, and Hitachi Ltd.

University of South Carolina, Columbia
September 1994

Jeeva S. Anandan
John L. Safko

Contents

SECTION 1

After-Banquet Talk in Honor of Aharonov's 60th Birthday

Symposium on Fundamental Aspects of Quantum Theory
Columbia, South Carolina, December 10-12, 1992

A. Pines
University of California, Berkeley

Yakir Aharonov: From A to B

Following the dictates of David Mermin, I have prepared some spontaneous remarks:

Ladies and Gentiles,

You see before you a most reluctant after-dinner speaker. Someone once said that if you took all the after-dinner speakers and laid them head-to-toe at the equator, ... that would be a very good thing. In fact, some years ago, my friend Anatole Abragam warned me – Alex, when they start asking you to give after-dinner speeches, it might be an indication that you are no longer on the way up. So when I was asked to talk about Aharonov tonight, the first two words that came to my mind were – oy vey.

But, ladies and gentlemen, this is no ordinary occasion – Yakir Aharonov is not only a truly great scientist and one of the most brilliant and stimulating people I have ever known, he is an extraordinary colleague and dear friend, and it is a privilege and a pleasure for me to say a few words about him. You might well ask, why me, a chemist, talking about a physicist. Well, Aharonov himself once paid me what he considers the greatest compliment you guys can give a chemist – Come on, Alex, you're not really a chemist, you're too smart, ... you're a physicist. Yakir, it's your birthday, let me return the compliment – you don't look seventy.

Yakir Aharonov was born in 1932, in Haifa, Israel, to Russian parents. He grew up, so to speak, in Kiryat Haim, where, already at age five, it was abundantly clear that he was a mathematical prodigy. The residents of Kiryat Haim soon became accustomed to the apparition of the boy Aharonov accosting and threatening them in the streets, challenging them to give him a problem -- a novel concept of mathematical mugging, your problem or your life.

Because his parents were unwilling to teach him chess (a waste of time), Aharonov traded some strawberries from his yard to a neighbor, an older child, who taught him the game. When not playing with his friend, Aharonov would play by himself, one hand against the other, one playing white and the other black. It is not known which hand was stronger, his left hand or his other left hand. As many of you know, Aharonov had a natural aptitude for the game and became a very strong player, today an Israeli candidate master. During his period as Miller Professor at Berkeley, Aharonov made an unforgettable impression not only on the scientists, but also on the nationally renowned Berkeley chess community. As a young man, Aharonov had a gift not only for math and chess; he was good at all sorts of games and puzzles. He discovered, to his joy, that his prowess at backgammon made him almost irresistible to middle-eastern women.

The last time I played blitz chess against Aharonov, he again asked if I wanted a handicap. I related to him the (perhaps apocryphal) story told to me recently by John Rowlinson about Max Euwe, the

former world chess champion. Euwe was on a train analyzing a game on his pocket chess set. A fellow traveler in the compartment asked him if he played chess, to which Euwe replied that yes, he did. Would you like to play a game, asked the other fellow; sure, said Euwe, who proceeded to set up the pieces and then removed one of his rooks. What are you doing, asked his partner. I'm giving you a rook, replied Euwe. You're giving me a rook? You've never played against me, you don't know who I am, how can you give me a rook? If I couldn't give you a rook, said Euwe, I'd know who you are.

Well, Aharonov doesn't give me a rook, but he does give me a differential time handicap in order to imbue the game with some semblance of balance. In other words, he beats the hell out of me. It is because of Aharonov that I have now resorted to playing for money against small children. But Aharonov too is fallible – about twenty five years ago, in New York, he played, and lost, three games against Bobby Fischer. Aharonov maintains that this is pretty good; he lost only three games, so he did better than the famous Russian, Taimanov, and the great Dane, Larsen, who each lost six games against Fischer.

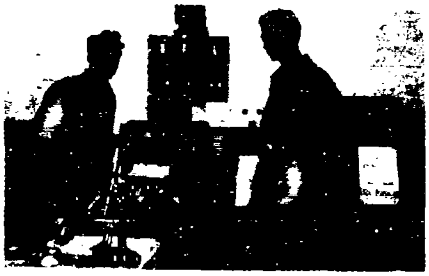
At age eleven, Aharonov took up the violin, an instrument that he cherishes to this very day. He soon discovered that the best acoustics for his instrument were in the kitchen and bathroom. It was later, after he read how Einstein had independently made the same discovery, that Aharonov decided he would become a physicist.

After graduation from high school, Aharonov was inducted into the army, into the artillery division. Yes, the artillery division. He soon lost interest in experimental artillery after he proved that quantum corrections to the ballistic trajectories were insignificant and, much to the relief of the commanding authorities, he volunteered for an army research unit. The only legacy of

Aharonov's army experience was his occasional, misguided tendency to force himself upon his friends as a bodyguard.

After his discharge from the army, Aharonov studied at the Technion, the Israel Institute of Technology, where he met the late David Bohm. Here Aharonov is shown at the Technion with a co-student whom he identifies as Tsachi Gozani. Gozani allegedly spent much of his time begging Aharonov to stay away from the apparatus. After discussions with faculty members who feared for their lives, Aharonov seriously contemplated becoming a theoretician. He moved with Bohm to Bristol to do his Ph.D. and it was there that the famous Aharonov-Bohm effect was conceived, elucidated and published.

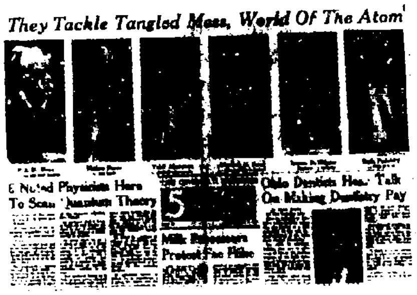
One of the external examiners for Aharonov's Ph.D. was Rudolph Peierls, who claimed he did not believe some argument that Aharonov had formulated about energy-time uncertainty, but Peierls could not find an error. He invited Aharonov to Birmingham, where they sat and argued for days, after which Peierls was convinced and said that he now believed. But Yakir tells me that just two years ago, Peierls was in Israel for the Landau Symposium - he ran into Aharonov and said hey, aren't you Aharonov? Yes, I am. Well, said Peierls, now I don't believe you again.



It was during his time in England that Aharonov became concerned about his Israeli accent, because he felt that it was hindering his chances with women. He arranged for intensive tutoring sessions in elocution, seeking to acquire not just any old accent, but an Oxford accent, and devoting considerable time and effort to the enterprise. On the day of the first experiment with his new accent, an excited Aharonov ventured into the streets of Bristol and asked for directions to go somewhere; I imagine that we can all sympathize with his frustration when the answer came back in Hebrew.

Following his Ph.D., Aharonov spent several years at Brandeis and Yeshiva Universities in the United States. In 1962, he created a sensation when he talked about the Aharonov-Bohm effect at the Cincinnati Conference on quantum theory (the other participants included Dirac, Furry, Podolsky, Rosen and Wigner). The conference made headlines despite the many other exciting events in Cincinnati at the time.

In 1966, Aharonov joined the faculty at South Carolina and, in 1967, he became Full Professor at Tel-Aviv University. He was subsequently honored with chairs in physics both at Tel-Aviv and here in South Carolina, where, I understand, he is again contemplating changing his accent. His colleagues here know that, for Aharonov, physics is not just a job – it is a passion, like chess. That Tel Aviv University and the University of South Carolina pay him to indulge in his passion remains for him unfathomable. Yakir, may it become yet more unfathomable.



They Tackle Tangled Mess, World Of The Atom

6 Noted Physicists Here To Scan Quantum Theory

Ohio Dentists Hear Talk On Making Dentistry Pay

Milk Producers Protest Fee Hike

Over the years, Aharonov has further cultivated, carefully and successfully, his image as a shlemiel, thereby shielding him from annoying appeals to help around the lab, the department or the house, and leaving him time to do what he loves and does best – to think. And, as many of us know, Aharonov thinks best in an atmosphere composed of ten percent oxygen, forty percent nitrogen and fifty percent cigar smoke. What kind of cigar smoke? Well, let's just say that many years ago, I gave him one of my prized Montecristos from Havana, and he was able to exchange it for a year's supply of his beloved White Owls. Aharonov continues with his tradition of visiting Berkeley whenever he runs out of cigars, much to the delight of my children, by whom he is much admired.

Yakir Aharonov is a giant of modern physics. From his Ph.D. with Bohm to his work on geometric phases, he has made monumental contributions to quantum theory, and he has profoundly advanced our understanding of electromagnetism and other gauge theories of fundamental interactions. On two occasions, John Maddox, the editor of Nature (the science magazine), suggested, justifiably, many of us believed, that Aharonov, Bohm and Berry should get the Nobel Prize for physics. In his first editorial on the subject, in 1989, Maddox writes about Abrahamov and the Abrahamov-Bohm effect; in his second editorial on the subject, this year, he makes a slightly better approximation, writing about Aharanov and the Aharanov-Bohm effect. And listen to the perverse, yet quaint 1989 description of the effect – Abrahamov and Bohm, independently of M. J. Berry, have shown that the supposedly insignificant complex phase of Maxwell's electromagnetic potential is measurable.

Well, Yakir Aharonov is no stranger to honor and to ceremony. He is a member of the Israel and U.S. National Academies of Sciences, and amongst his many awards are the prestigious Israel Prize in exact sciences and the Elliot Cresson Medal of the Franklin Institute in Philadelphia. But Aharonov is particularly proud of the knighthood bestowed upon him by his friends on the occasion of his fiftieth birthday which, he calculates, was ten years ago. I guess the citation reads – why is this knight different from all other knights?

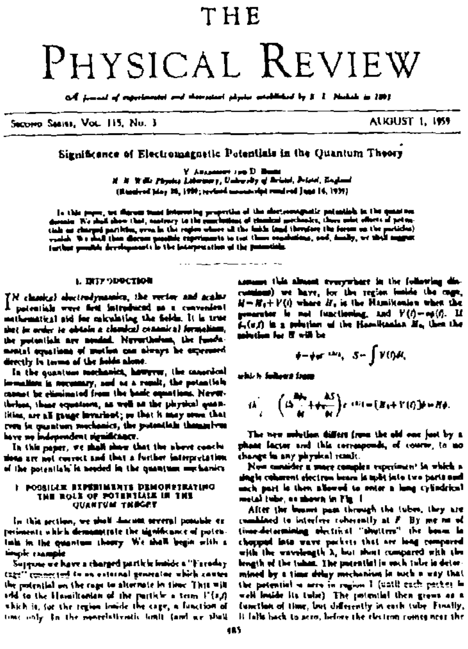
Ladies and gentlemen, I was asked to make my remarks either wi'ty or brief – so I must come to a close.

Yakir Aharcnov is a man with a legendary hunger for science and for life. But beyond his genius and his accomplishments, Aharonov has that rarest of human qualities – he is a mensch. Dear Yakir, I am sure that I speak on behalf of everyone here when I say that you have earned our respect. On the occasion of your sixtieth birthday, permit me to offer a toast to you and your family – the Aharonovs, the Abrahamovs and the Aharanovs – Yakir and Nilli,......... to another sixty years.

פיזיקה של עולם קפריזי

הכל בראש

[illegible]



THE

PHYSICAL REVIEW

A journal of experimental and theoretical physics established by E. L. Nichols in 1893

SECOND SERIES, VOL. 115, No. 3 AUGUST 1, 1959

Significance of Electromagnetic Potentials in the Quantum Theory*

Y. AHARONOV AND D. BOHM
H. H. Wills Physics Laboratory, University of Bristol, Bristol, England
(Received May 28, 1959; revised manuscript received June 16, 1959)

In this paper, we discuss some interesting properties of the electromagnetic potentials in the quantum domain. We shall show that, contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish. We shall then discuss possible experiments to test these conclusions, and, finally, we shall suggest further possible developments in the interpretation of the potentials.

1. INTRODUCTION

IN classical electrodynamics, the vector and scalar potentials were first introduced as a convenient mathematical aid for calculating the fields. It is true that in order to obtain a classical canonical formalism, the potentials are needed. Nevertheless, the fundamental equations of motion can always be expressed directly in terms of the fields alone.

In the quantum mechanics, however, the canonical formalism is necessary, and as a result, the potentials cannot be eliminated from the basic equations. Nevertheless, these equations, as well as the physical quantities, are all gauge invariant; so that it may seem that even in quantum mechanics, the potentials themselves have no independent significance.

In this paper, we shall show that the above conclusions are not correct and that a further interpretation of the potentials is needed in the quantum mechanics.

2. POSSIBLE EXPERIMENTS DEMONSTRATING THE ROLE OF POTENTIALS IN THE QUANTUM THEORY

In this section, we shall discuss several possible experiments which demonstrate the significance of potentials in the quantum theory. We shall begin with a simple example.

Suppose we have a charged particle inside a "Faraday cage" connected to an external generator which causes the potential on the cage to alternate in time. This will add to the Hamiltonian of the particle a term $V(x,t)$ which is, for the region inside the cage, a function of time only. In the nonrelativistic limit (and we shall assume this almost everywhere in the following discussions) we have, for the region inside the cage, $H=H_0+V(t)$ where H_0 is the Hamiltonian when the generator is not functioning, and $V(t)=e\varphi(t)$. If $\psi_0(x,t)$ is a solution of the Hamiltonian H_0, then the solution for H will be

$$\psi=\psi_0 e^{-iS/\hbar},\quad S=\int V(t)dt,$$

which follows from

$$i\hbar\frac{\partial\psi}{\partial t}=\left(i\hbar\frac{\partial\psi_0}{\partial t}+\psi_0\frac{\partial S}{\partial t}\right)e^{-iS/\hbar}=[H_0+V(t)]\psi=H\psi.$$

The new solution differs from the old one just by a phase factor and this corresponds, of course, to no change in any physical result.

Now consider a more complex experiment in which a single coherent electron beam is split into two parts and each part is then allowed to enter a long cylindrical metal tube, as shown in Fig. 1.

After the beams pass through the tubes, they are combined to interfere coherently at F. By means of time-determining electrical "shutters" the beam is chopped into wave packets that are long compared with the wavelength λ, but short compared with the length of the tubes. The potential in each tube is determined by a time delay mechanism in such a way that the potential is zero in region I (until each packet is well inside its tube). The potential then grows as a function of time, but differently in each tube. Finally, it falls back to zero, before the electron comes near the

485

SECTION 2

AHARONOV–BOHM EFFECT AND GEOMETRIC PHASES

DYNAMIC OBSERVATION OF FLUX LINES BASED ON THE AB EFFECT PRINCIPLE

A. TONOMURA
Advanced Research Laboratory, Hitachi, Ltd.
&
Tonomura Electron Wavefront Project, ERATO, JRDC
Hatoyama, Saitama 350-03, Japan

ABSTRACT

Flux lines penetrating superconducting films are directly observed with a "coherent" field-emission electron beam. These flux lines are detected as phase shifts of an electron beam passing through the films due to the Aharonov-Bohm effect.

1. INTRODUCTION

The behavior of flux lines plays a decisive role in the fundamentals and practical applications of superconductivity.

Although much effort has been expended on developing methods to directly observe flux lines, until recently flux lines have evaded direct observation because they are shaped like an extremely thin thread and have a small flux value of $h/2e(= 2 \times 10^{-15}$ Wb). In 1967, Essman and Träuble[1] used the Bitter technique to directly observe the flux-line lattice predicted by Abrikosov.[2] In this technique, fine ferromagnetic particles are sprinkled over the superconductor surface and the location of flux lines is observed as a replica with an electron microscope. This technique has recently been used to elucidate the microscopic characteristics of high-Tc superconductors.[3] However, this technique is essentially static, and it cannot determine the dynamic behavior of flux lines. New techniques for observing flux-lines have also been developed.[4,5] For example, Hess, *et al.*[4] used a scanning tunneling microscope to observe the flux-line lattice of $NbSe_2$. However, dynamic observation is still not feasible with these techniques.

The possibility of direct observation using a transmission electron microscope has been theoretically investigated making use of the fact that an electron beam is deflected[6-8], or phase-shifted by flux lines: Despite trials, the deflection angle is too small——less than 1×10^{-6} rad——to observe flux lines as a Lorentz micrograph (a greatly defocused electron micrograph). Or, in other words, the phase shift of the electron beam is produced due to the Aharonov-Bohm effect[9] when the beam passes through a flux line, which is, in this case, less than π. This phase shift was actually detected by electron interferometry.[10,11] Using this method, Boersch, et al.[12] observed the location of a single flux line leaking from a superconducting tube as a shift of parallel interference fringes by half their spacing, followed by thermally activated jumps of flux lines from one pinning center to another with a time resolution of around one second.

However, as a result of the development of a "coherent" field-emission electron beam,[13,14] it has become possible to measure the phase distribution of an electron beam to a precision of 1/100 of the wavelength[15] through electron holography.[16-18] In addition, the two-beam interference pattern has become directly observable on the fluorescent screen, permitting dynamic observation.

Such technical development has helped to open the way to direct observation of flux lines. In this method, a single flux line leaking from a superconductor surface could be observed directly and even dynamically as a contour fringe in an interference micrograph. Furthermore, for the first time, flux lines were also observed in the transmission mode.

2. EXPERIMENTAL APPARATUS

Experiments were carried out using holography electron microscopes. The holography electron microscopes used in the present experiments are transmission electron microscopes equipped with field-emission electron guns[13,14] for coherent specimen illumination, and electron biprisms[19] for hologram formation.

A cut-away drawing of our 350-kV holography electron microscope[20] is shown in Fig. 1. The main column below the objective lens is almost the same as that of a Hitachi H-9000 transmission electron microscope. The illumination system consists of a cold field-emission electron gun and double condenser lenses.

The specimen is illuminated by a collimated electron beam. The small illumination angle 2β, which is indispensable for forming electron holograms, or Lorentz micrographs, can be reduced to 5×10^{-8}

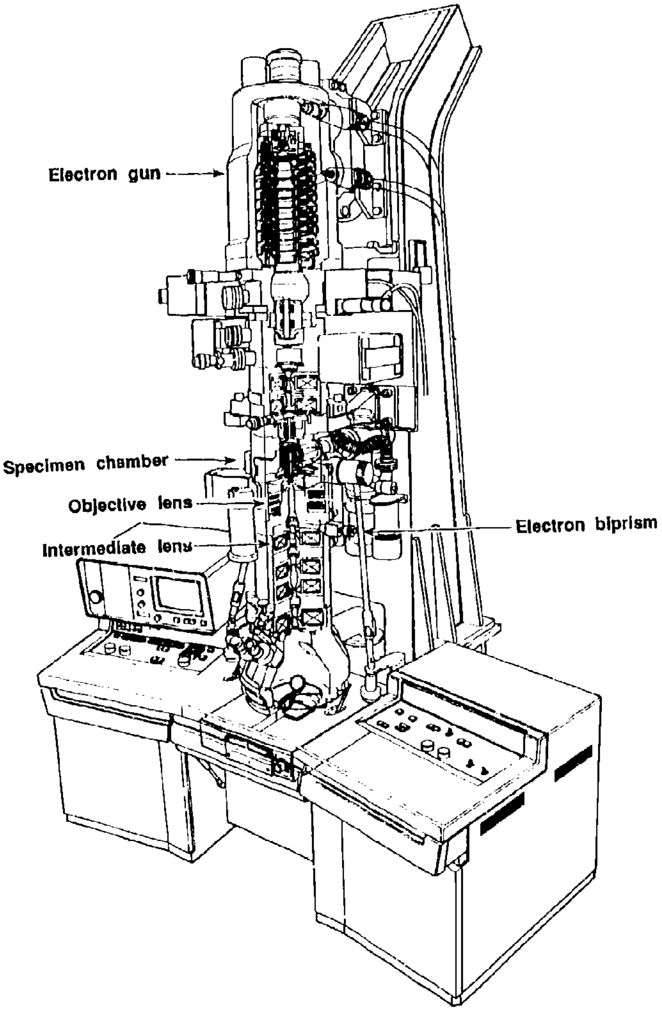


Fig. 1. 350kV holography electron microscope.

rad by the double condenser lenses. A low-temperature specimen stage is substituted when flux lines re observed. The temperature of the specimen holder can be reduced to 4.5K. At the same time, a magnetic field of up to 150 Gauss can be applied in the horizontal direction.

Electron biprisms[19] are installed at two positions, one below the objective lens and the other below the intermediate lens. The appropriate biprism position can be selected after the optical conditions such as magnification have been determined.

The specimen or hologram image can be enlarged by magnifying lenses as in the electron microscope, which is usually recorded on film. However, for dynamic observation, it is recorded on videotape through a television system attached to the microscope.

3. EXPERIMENTAL METHOD

Two methods were employed in the present experiments, i.e., electron holography and Lorentz microscopy. Magnetic lines of force leaking from the superconductor surface were directly observed as contour fringes in an electron interference micrograph obtained through the electron holography process. In the Lorentz micrograph with an appropriate defocusing, flux lines in the superconductor were observed as globules with black and white contrast pairs.

3.1 Electron Holography

Electron holography [16] is a two-step imaging method using electron waves and light waves (see Fig. 2). An electron wave illuminates an object and is scattered. A reference wave that has been tilted by a prism is then projected onto the scattered wave to form an interference pattern that is recorded on film. This film, called a hologram, is subsequently illuminated by a collimated laser beam. The exact image is then three-dimensionally reproduced. An additional conjugate image is also produced in holography.

Once electron wavefronts have been reproduced as light wavefronts, versatile optical techniques can be used to supplement electron optics.

An interference micrograph, or contour map of the wavefront, can be obtained by simply overlapping an optical plane wave with this reconstructed wave (see Fig. 3(a)). If a conjugate wavefront instead of a plane wave overlaps this wavefront, the phase difference becomes twice as large, and is as if the phase distribution were amplified two times, as shown in Fig. 3(b). By repeating this technique, a phase shift can be detected even as small as 1/100 of a wavelength.

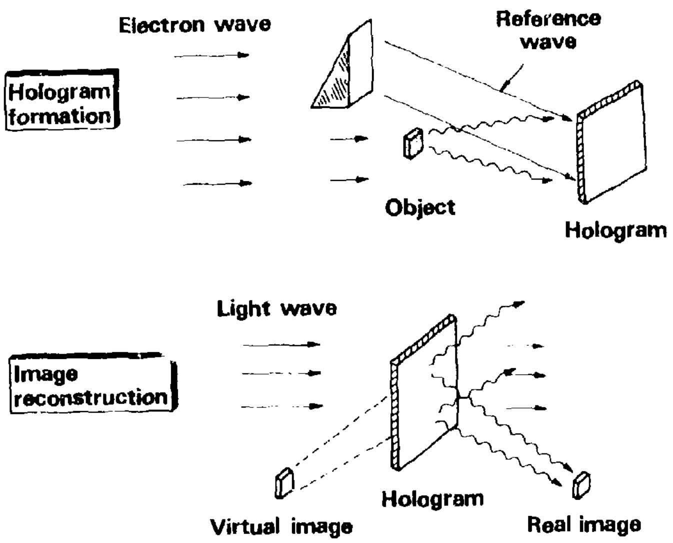


Fig.2. Principle behind electron holograpy.

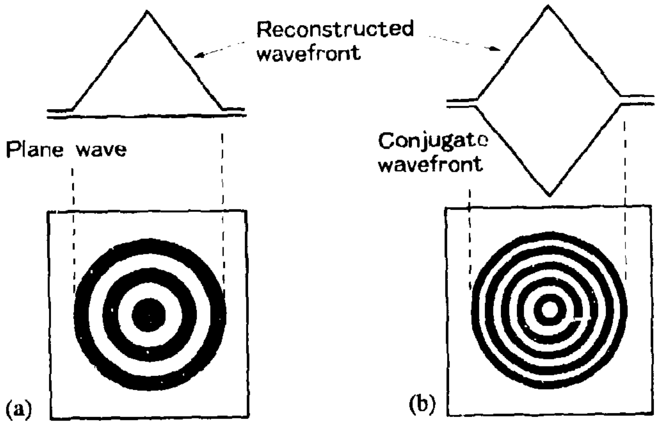


Fig.3. Principle behind phase amplification.
(a) Contour map. (b)Twice-amplified contour map.

This phase-amplified interference electron microscopy provides information about microscopic distribution of the electric[21] and magnetic [22] fields.

Flux lines can be directly observed in a twice phase-amplified interference micrograph. The observation principle is illustrated in Fig. 4. When an electron beam is incident to a uniform magnetic field, the beam is deflected to the left by the Lorentz force, which acts perpendicularly to the direction of the magnetic field. Viewing electrons as waves, the introduction of a "wavefront" perpendicular to the electron trajectory will suffice. The incident electron beam is a plane wave, but the outgoing beam becomes a plane wave with the left side tilted up. In other words, the wavefront is viewed as rotating around a rotating axis; the magnetic line of force. From a contour map of this wavefront, it can be seen that the contour lines follow the magnetic lines of force. This is because the height of the magnetic lin: of force is the same along it. Thus, a very simple conclusion can be reached: when a magnetic field is observed in an interference electron micrograph, the contour fringes can be considered to represent magnetic lines of force.

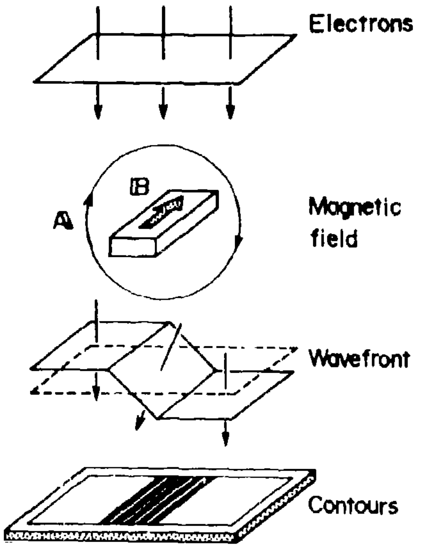


Fig.4. Principle behind magnetic flux observation.

The interference fringes are also quantitative. A simple calculation convinces us that a certain minute amount of magnetic flux, h/e, is flowing between adjacent contour fringes. This is, in a sense, quite natural. A superconductive flux meter, SQUID, can measure the flux in units of $h/2e$ by using Cooper pair interference. An electron interference micrograph is formed by the interference of electrons rather than Cooper pairs. Therefore, the flux unit is h/e, since the electric charge is e rather than $2e$. However, the principle is the same.

It can be concluded then that a contour fringe in a twice phase-amplified interference micrograph indicates a single flux line.

3.1.1 Lorentz microscopy

A Lorentz micrograph is a greatly defocused electron micrograph. The principle behind it is shown in Fig. 5. When an electron beam is incident to a ferromagnetic thin film, which has two magnetic domains, the beam is deflected by the magnetization, and the deflection directions are different for th^ two domains. Therefore, when the elec

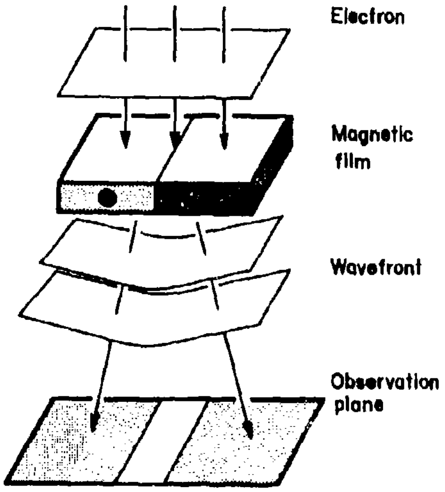


Fig.5. Lorentz microscopy.

tron intensity distribution is observed in the lower plane, the domain wall can be observed as a line of the weak intensity. Thus, Lorentz microscopy is effective when the magnetic field changes suddenly, such as at a domain boundary in a ferromagnetic film. However, it is not easy to observe flux lines in free space by Lorentz microscopy, since magnetic fields there are distributed smoothly in a harmonic form.

4. EXPERIMENTAL RESULTS

Individual flux lines were statically and dynamically observed using holography electron microscopes.

4.1 Observation in the Profile Mode

Flux lines leaking out from a superconductor surface can be directly observed as contour frin s in a twice phase-amplified interference micrograph through electr n holography, as explained in the previous section.

The experimental arrangement is shown in Fig. 6. A thin tungsten wire 40μm in diameter was used as the substrate for a superconducting specimen. Lead was evaporated onto one side of the wire. A magnetic field of a few Gauss or less was applied to the evaporated lead film. The specimen was cooled to 4.5K. In a weak magnetic field, the magnetic lines are excluded from the superconductor by the Meissner effect,

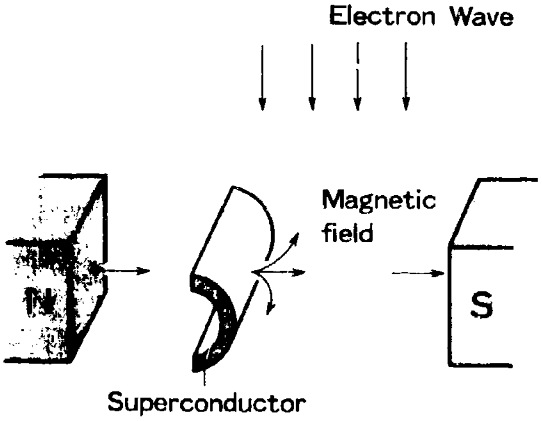


Fig.6. Experimental arrangement to observe flux lines

but if the magnetic field is strong, the magnetic lines of force penetrate the superconductor in the form of flux lines. By applying an electron beam to the specimen from above, the magnetic lines of force of flux lines were observed through the process of electron holography.

Figure 7(a) shows the single flux line observed when the superconducting film was 0.2μm thick. In this figure, the phase difference is amplified by a factor of two. Therefore, one interference fringe corresponds to one flux line. A single flux line is captured in the right part of this photograph. The magnetic line of force is produced from an extremely small area of the lead surface, and then spreads out into free space.

In addition to observing isolated flux lines, another surprising result was found. A pair of flux lines were observed that were oriented in opposite directions and connected by magnetic lines of force (Fig. 7 (a) left). The following explanation may be considered. When the specimen is cooled below the critical temperature, the lead becomes superconductive. During the cooling, however, the specimen experiences a state where the flux-line pair appears and disappears repeatedly due to thermal excitation [23] and is pinned by some imperfection in the superconductor, eventually resulting in the flux being frozen.

What happens when the thickness of the superconducting thin film is increased? Figure 7(b) shows the state of the magnetic lines of force when the thickness is 1μm. It can be seen that the state changes completely. Magnetic flux penetrates the superconductor not as in-

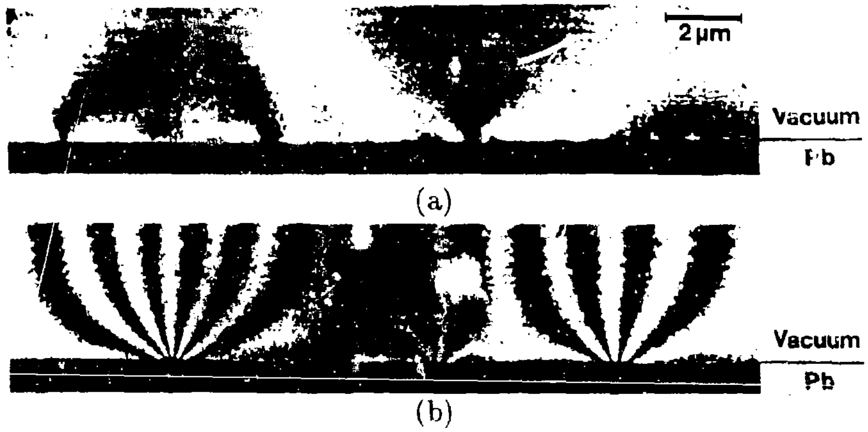


Fig.7. Interference micrograph of flux lines leaking from Pb film (Phase amplification:×2).
(a)Thickness = 0.2μm. (b)Thickness = 1μm.

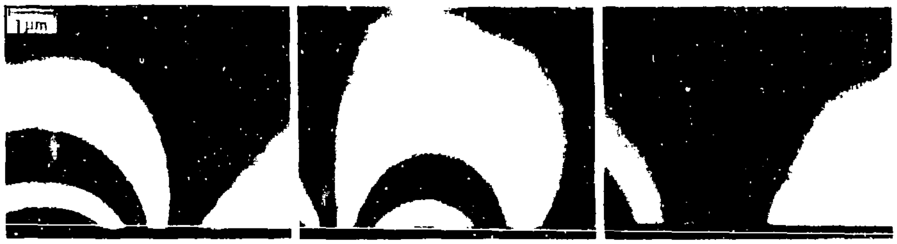
dividual flux lines but in a bundle. The figure does not show any flux-line pairs.

Our explanation for this phenomenon is as follows. Becauce lead is a type-I superconductor, the strong magnetic field applied to it partially destroys the superconductive state in some parts of the specimen (intermediate state). Figure 7(b) is a photograph showing that the magnetic lines of force penetrate the parts of the specimen where superconductivity has been destroyed. However, since the other surrounding parts are still superconductive, the total amount of penetrating magnetic flux is an integral multiple of the flux quantum, $h/2e$. Thin superconducting films (Fig. 7(a)) were an exception. In that case, however, lead behaved like a type-II superconductor and the flux penetrated the superconductor in the form of individual flux lines.

Since the flux itself can be observed using electron holography, its dynamic behavior can be observed.[24] In this case, after electron holograms were dynamically recorded on videotape, a twice phase-amplified contour map of each frame was numerically reconstructed, and again recorded on videotape. Although off line, flux dynamics could be observed with a time resolution of 1/30 of a second.

The experiment was carried out as follows. Trapped fluxes in a Pb thin film remained stationary at 5K. However, when the sample temperature was raised, the flux line diameter gradually increased. Just below the critical temperature, the flux lines began to move. Figure 8 shows a section from the videotape that recorded this movement.

Three flux lines in the upward direction are trapped in the superconductor and their magnetic lines of force can be seen in Fig. 8(a). At 0.13 seconds, the flux lines moved suddenly to the left after only the lapse of a single frame. Two upward flux lines and two downward



(a) (b) (c)

Fig.8. Dynamical observation of trapped flux line near Tc.

(a) 0 seconds. (b)0.13 seconds later. (c) 1.33 seconds later.

flux lines are connected by magnetic lines. At 1.33 seconds, downward flux lines moved to the right and only a broad single magnetic line remained.

Although this flux movement due to thermal activation is random, a similar experiment is now in progress where a current is applied to the superconductor. In this case, flux lines receive a Lorentz force determined by the current, but with opposite directions for upward and downward flux lines. The pinning force at each pinning site can thus be measured.

4.2 Observation in the Transmission Mode

Flux lines have recently been observed in the transmission mode.[25) A two-dimensional distribution of flux lines was seen dynamically by Lorentz microscopy with a 300-kV holography electron microscope.

The experimental arrangement is shown in Fig. 9. A Nb thin film was prepared by chemically etching a roll film. The film, set on a low-temperature stage, was tilted at 45° to the incident beam with 300 keV electrons falling vertically, so that the electrons could receive the flux-line magnetic fields penetrating the sample perpendicularly to its surface. An external magnetic field of up to 150 Gauss was applied horizontally.

The information about the flux lines is contained in the phase distribution, or in other words, the wavefront distortion of the transmitted electron beam. This information cannot be read from a conventional electron micrograph where only the intensity is recorded. However, the distortion reveals itself in a defocused image, *i.e.*, a Lorentz micrograph, in which a flux line can be seen as a tiny spot; one half bright and the other half dark.

The sample was first cooled down to 4.5K and the applied magnetic field B was gradually increased. As B was increased, flux lines suddenly began to penetrate the film at $B = 32$ Gauss, and their number increased with B. Their dynamic behavior was quite interesting: at first, only a few flux lines appeared here and there in the field of view, $15 \times 10\mu m^2$, oscillating around their own pinning centers and occasionally hopping from one center to another. These movements continued as long as the flux lines were not closely packed ($B \leq 100$ Gauss).

An example of the equilibrium Lorentz micrographs at $B = 100$ Gauss is shown in Fig. 10. The film has a fairly uniform thickness in the region shown, but is bent along the black curves, called bend contours, which are due to Bragg reflections at the atomic plane brought to a favorable angle by bending. Each spot with a black and white contrast is the image of a single flux line. This contrast reversed, as expected, when the applied magnetic field was reversed. The tilt direction of the sample can be read from the line dividing the black and white part of the spots. Since the black part is on the same side for all the spots, the polarities of all the flux lines as seen in the region aie the same.

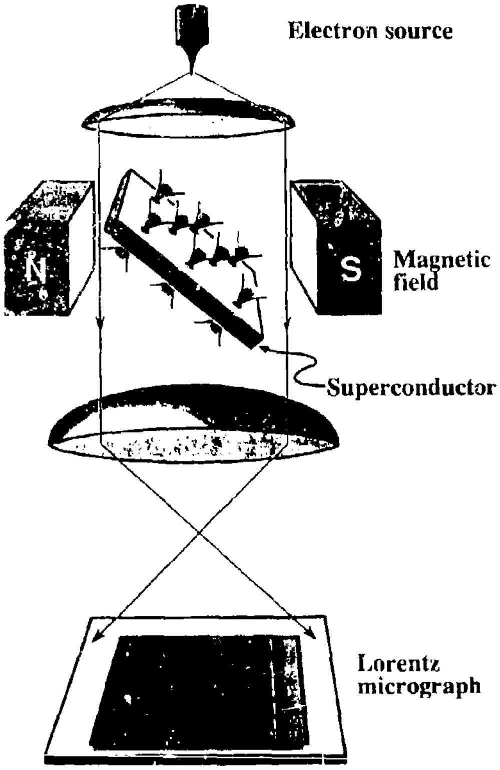


Fig.9. Schematic for flux-line lattice observation.

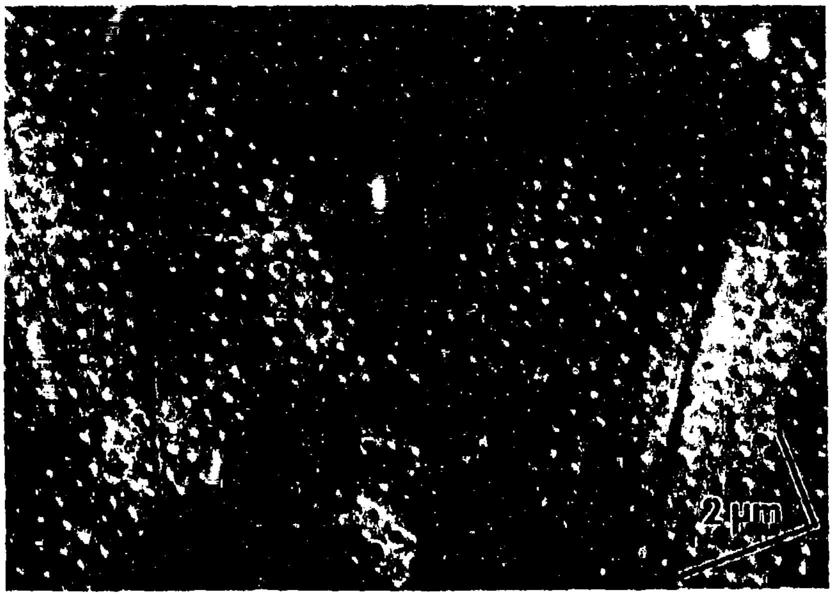


Fig.10. A Lorentz micrograph of a two-dimensional array of flux lines in superconducting Nb film.

At low B, *i.e.*, up to 30 - 50 Gauss, the flux lines are too scarce to form a lattice, even in equilibrium. At B = 100 Gauss where the flux-line density is so high that it cannot be anything but a hexagonal lattice, the flux-line configuration and movement are influenced by structure defects.

5. CONCLUSION

Electron holography has opened up a new window for direct and real-time observation of the microscopic dynamics of individual superconducting flux lines such as in flux creep, pinning, etc, which up to now has only been observed in macroscopic experiments. This technique will effectively be employed for elucidating fundamentals and practical applications of superconductivity, especially in the field of high-Tc superconductors.

6. REFERENCES

1) V. Essman and H. Träuble: *Phys. Lett.* **A24** (1967) 526.
2) A. A. Abrikosov, *Zh. Eksp. Teor. Fiz.* **32** (1957) 1442 [Sov. Phys. JETP 5 (1957) 1174].
3) For example, D. J. Bishop, et al.: *Science* **255** (10 January 1992) 165.
4) H. F. Hess, et al.: *Phys. Rev. Lett.* **62** (1989) 214.
5) J. Mannhart, et al.: *Phys. Rev.* **B35** (1987) 5267.
6) H. Yoshioka: *J. Phys. Soc. Jpn.* **21** (1966) 948.
7) C. Colliex: *Acta Crystallogr. Sect.* **A24** (1968) 692.
8) C. Capiluppi, G. Pozzi, and V. Valdre: *Phil. Mag.* **26** (1972)865.
9) Y. Aharonov and D. Bohm: *Phys. Rev.* **115** (1959) 485.
10) H. Wahl: *Optik* **28** (1968) 417.
11) B. Lischke: *Phys. Rev. Lett.* **22** (1969) 1366.
12) H. Boersch, et al.: *Phys. Status Solidi* (b) **61** (1974) 215.
13) A. V. Crewe, et al.: *Rev. Scient. Instrum.* **39** (1968) 576.
14) A. Tonomura, et al.: *J. Electron Microsc.* **28** (1979) 1.
15) A. Tonomura, et al.: *Phys. Rev. Lett.* **54** (1985) 60.
16) D. Gabor: Proc. *R. Soc. London*, Ser. **A197** (1949) 454.
17) A. Tonomura: *Physics Today* **22** (April 1990) 22.
18) A. Tonomura: *Adv. Phys.* **41** (1992) 59.
19) G. Möllenstedt and H. Düker: *Z. Physik* **145** (1954) 377.
20) T. Kawasaki, et al.: *Jpn. J. Appl. Phys.* **29** (1990) L508.
21) S. Frabboni, G. Matteucci & G. Pozzi, *Phys. Rev. Lett.* **55** (1985) 2196.
22) A. Tonomura, et al., *Phys. Rev. Lett.* **44** (1980) 1430.
23) J. M. Kosterlitz and D. Thouless, *J. Phys.* **C6** (1973) 1181.
24) T. Matsuda, et al., *Phys. Rev. Lett.* **66** (1991) 457.
25) K. Harada, et al.: *Nature* **360** (5 November 1992) 51.

SIGNS AND MIRACLES OF THE AHARONOV-BOHM EFFECT

Alfred S. Goldhaber
Institute for Theoretical Physics
State University of New York
Stony Brook, NY 11794-3840

ABSTRACT

Familiar aspects of electromagnetic influences on the quantum propagation of charged particles – and some not so familiar – conspire to support a view of the Aharonov-Bohm effect as the essential and primary manifestation of gauge interactions. In particular, the perturbative renormalization group scaling for this form of the coupling lends appeal to the notion that, on scales where the conventional coupling α becomes strong, there should be a 'universal pasta solution' for the vacuum structure of any gauge theory: The Nielsen-Olesen proposal of flux spaghetti should apply not only for QCD at long distances as they argued, but just as well for QED at short distances.

Signs of the AB effects

I hope to weave into a single tapestry a number of threads which together illustrate the beauty as well as the power of the Aharonov-Bohm effect[1] as an organizing principle for gauge theories. Some of these notions are explicit, some perhaps implicit in the existing literature. Much of the analysis is contained in a recent paper with Hsiang-Nan Li at Academia Sinica in Taiwan and Rajesh Parwani at Saclay,[2] and I am most grateful to them for a stimulating, still progressing collaboration. Let me begin by addressing a deceptively simple question, "What is the sign of the AB effect?" I failed to grasp the point properly in my spoken presentation, but Jeeva Anandan and Raymond Chiao helped me afterwards to see that the AB effect really is two complementary effects: There is the shift of interference fringes which Aharonov and Bohm pointed out in their original work,[1] and then there is the shift in angular momentum eigenvalue for a particle in a ring encircling some magnetic flux.

Let us start by determining the sign of the second effect, the shift in angular momentum eigenvalue. This may be done by classical physics using Ehrenfest's theorem, which states that the change in expectation value of some observable is determined by the classical equation of motion for that observable, with the appropriate expectation value used to compute the classical force. Imagine that the magnetic flux is turned on adiabatically, so that the particle remains in a definite eigenstate throughout. By Faraday's law, if the flux is generated by a current of particles with the same sign of charge as the test particle, then the angular momentum of the test particle must decrease as the angular momentum of the current particles increases. Thus, for positive charge-flux product $q\Phi$, with the flux coming out of the plane of motion as viewed from above, the shift in angular momentum is

$$\delta M = -q\Phi/hc = -F\ , \qquad (1)$$

where F is the flux expressed in units of an AB quantum, hc/q. In terms of signs, this means that the sign of the coherent-ring AB effect is negative.

Next we need to study the classic fringe-shift effect. To put the question in terms of observables, let us ask: On which side of a flux, the right or the left, should one introduce an attractive, velocity-increasing electrostatic potential in order to compensate the AB phase? To answer this question by classical physics, consider a charged particle traveling through a region of uniform magnetic field oriented up with respect to the plane of motion. How could we arrange that the particle travels in a straight line instead of being deflected to the right? We could compensate for the Lorentz force by introducing an electric field in the plane, which by itself would push the particle to the left. This means that the electrostatic potential decreases from right to left, and thus is more negative on the left than on the right.

Since a uniform magnetic field may be described as a collection of adjacent regions of magnetic flux, it follows that to compensate for the AB phase one must place a suitable negative potential on the left side of the flux. We can see this directly:[3] The pure uniform magnetic field gives a deflection to the right. This is the same effect which would result if the phase velocity were increased on the right, since that means the number of wavelengths per unit distance increases, or the wavelength shortens, which by standard refraction ideas gives deflection to the right. To compensate then requires adding attraction on the left. Evidently this means that another observable, the direction of shift in the interference pattern, also must be to the right, so that the wave fringe motion of the AB effect is in the same direction as the classical deflection by a uniform field. The conclusion is that with standard conventions the sign of the classic AB effect is positive.

To see why these two opposite signs not only are compatible but are intrinsically connected, let us go to a special gauge, in which outside the region of magnetic field the vector potential vanishes almost everywhere, but between the azimuthal angles $\phi = 2\pi - \epsilon$ and $\phi = 0 + \epsilon$ there is a sharp jump in the phase of the wave function. Since the angular momentum is reduced by the flux, it follows that in this gauge the phase must have a decreasing contribution proportional to the flux as ϕ increases from 0 to 2π. Hence, the phase jump as ϕ increases through 2π must be positive, to restore the original value. What does this mean for interference shifts? If we imagine the right and left parts of the diffracted wave arriving at a distant screen at an angle $\phi > 0$, then the part of the wave which goes round on the right passes through the matching angle and experiences a positive phase jump. On the other hand, if we look on the screen at an angle $\phi < 2\pi$ then the wave which goes round on the left experiences a negative phase jump. In either case, the effect of the flux is to produce a positive relative phase shift of the right with respect to the left part, reproducing the previous conclusion.

Scattering on a thin flux string

Having learned the basic signs of the AB effect, we should look for miracles beyond the miracle of the effect itself. The first such miracle is found already in the original work.[1] Aharonov and Bohm observed that for a spinless charged particle interacting with an infinitely thin string of noninteger flux the AB effect is self-enforcing. In every partial wave, even the lowest, there is a centrifugal barrier which assures that the wave function vanishes as a positive power of a/λ, where a is the radius of the flux string, and λ is the De Broglie wavelength of the particle. Thus a low-energy particle effectively is excluded from the region of magnetic field, and this constitutes the necessary requirement that the sole observable consequence of the field is the AB effect.

I cannot resist an aside about the special case of nonzero integer F. For any integer F, there exists a partial wave which outside the flux has vanishing kinetic angular momentum, and which approaches a constant at small radius. For nonzero F this channel has a repulsive phase shift of order $1/ln(\lambda/a)$, implying that a/λ must be exponentially small if the phase shift is to be negligible. Thus at finite a there can be a significant correction to the limiting form valid for $a = 0$, and that correction violates periodicity because it distinguishes between $F = 0$ and all other integer values. This is not the last time that logarithmic effects will emerge in our consideration of flux strings.

The miracle of the thin string limit does not end with self-enforcement of the AB effect. If the ratio a/λ may be neglected then it is possible to compute the scattering amplitude analytically, and the result is remarkably simple. The amplitude is

$$f = \sin(\pi F)e^{-i\phi/2}/(2\pi ik)^{1/2}\sin(\phi/2) \ , \tag{2}$$

where k is the wave number of the charged particle.[1,4] I believe that for a suitable choice of gauge convention the above expression can be used in the F interval $[0, 1]$, with periodicity used to define the expression outside that interval, at the cost of a discontinuous derivative df/dF at integer values. The resulting cross section in any case is periodic in F with period 1, as all observables must be under these conditions.

Enter helicity

The situation changes in a significant way when the charged particle is an electron with the Dirac gyromagnetic ratio 2. Now the attractive interaction between the flux and the electron for parallel orientation of its magnetic moment allows penetration into the flux, and hence a sensitivity to more than the AB phase or the fractional part of F. What may be surprising is that in the long wavelength limit the sensitivity to F is only slightly enhanced: The observables depend not only on the fractional part $F-[F]$, but also on the sign $F/|F|$.[5] Thus, a new sign has entered the discussion, the sign of the magnetic flux. If electrons are confined to a cylinder centered on the flux, then energy levels in the partial wave with smallest kinetic angular momentum are lower for magnetic moment parallel to the flux than antiparallel. It is hard to decide which is more remarkable - the breakdown of periodicity in the dependence of observables on the flux, or the extremely simple form of that breakdown, leading to periodicity for nonnegative F, and separately for nonpositive F, but not for integer shifts which cross $F = 0$.

The effect on the scattering amplitude f of the magnetic moment coupling may be understood from a symmetry so powerful that it is fairly called miraculous, the conservation of helicity for an ideal Dirac electron in the presence of a pure magnetostatic field. The minimum modification required to bring the amplitude for scattering of spinless particles to a suitable form is the inclusion of a factor which rotates the spin in such a way that helicity eigenstates with respect to the initial beam direction are converted to eigenstates with respect to the scattered beam direction.[6] This factor is a spin rotation matrix $e^{i\sigma_3\phi/2}$, which however is not single-valued in ϕ. Since we are working in a gauge where the vector potential is nonsingular, the wave function and hence the scattering amplitude must be single-valued, and therefore we need to include a further factor $e^{\pm i\phi/2}$. The choice of sign for the exponent in this factor is directly related to observable quantities, the (opposite) signs of the phase shifts in the two partial waves with magnitude of kinetic angular momentum $J_3 = L_3 + s_3$ smaller than $1/2$. It turns out that the phase shift for Dirac magnetic moment parallel to the flux F is attractive, while for antiparallel it is repulsive. The scattering amplitude indeed is sensitive to the sign of the flux, and there are observable consequences, such as the Zeeman splittings mentioned above, and the properties of specially designed junctions.[7,2]

The case of nonzero integer F is altered a bit from the situation described earlier for spinless charged particles. The wave with orbital kinetic angular momentum zero and Dirac moment antiparallel to F again experiences a repulsive phase shift vanishing as $1/ln(\lambda/a)$, but there is no appreciable phase shift for the wave with Dirac moment parallel.

All the results of helicity conservation follow from the assumption that the electron experiences only magnetic forces. In many laboratory examples, this is not a good assumption, since the materials in conducting coils, and shields for those coils, exert powerful nonmagnetic forces on any incident particle. This difficulty may be overcome by making use of a purely magnetic field, as in the region just at the end of a tube containing a superconductor quantum of flux, $F = hc/2e$.[7] Another way to make the effective field purely magnetic is to deal only with propagation of electron quasiparticles through a superconductive medium, in which case the interaction is purely magnetic (and even locally a pure gauge effect) unless the quasiparticle actually penetrates a vortex of magnetic flux. In the interior of the vortex there might be an effective scalar potential influencing the motion. However, as long as the resulting forces are weak on the scale determined by the vortex radius they have negligible influence on long-wavelength scattering, so that helicity conservation continues to be a good approximation in this regime, even though no longer exact.

Such is the situation expected for cosmic strings. The ordinary vacuum plays the role of a superconducting medium, and for light fermions with effective mass coming from a Higgs coupling the change of effective mass in the interior of the string has negligible influence on the scattering. Thus, for cosmic strings one expects the helicity conserving boundary conditions, only changing the sign of one phase shift from that for pure AB scattering, to be the uniquely selected description for the effect on light, low-energy fermions of such strings.

Induction of vacuum currents

A further dynamical consequence of sensitivity to the sign of F may be seen if we examine (for spinor QED) vacuum electric currents induced by F. These currents always work to generate a magnetic field opposed to F, but otherwise periodic for nonnegative or nonpositive F.[8,9,10,2] In the case of scalar electrodynamics, the induced currents work to bring the flux to the nearest integer value, and so are insensitive to the sign and completely periodic in F.[11,8,2] A case for which one may guess the behavior, though it is not yet computed to my knowledge, is that of vector electrodynamics. Here perturbation theory, as well as studies of behavior of the vacuum in the presence of a uniform magnetic field, lead to the expectation that the induced currents will enhance rather than oppose the applied flux,[12,13,14] in other words, that conventional screening will be replaced by antiscreening. In all cases, one expects the induced flux to vanish when F is exactly an integer, since then the effective boundary conditions on any charged-particle wave function at the location of the infinitely thin flux are exactly the same as if no flux were present.[15] The antiscreening may be understood qualitatively because the attractive magnetic moment interaction reduces the effective mass of a spinor or vector particle. For spinors, the vacuum is described as a filled negative-energy sea, so that a reduction in effective mass actually raises the vacuum energy, while for vectors the reduction in single-particle energies implies also a reduction in the energies associated with the zero-point motion of the oscillator for each single-particle state ,[13,14,2] and hence a reduction in vacuum energy.

In the region close enough to the flux string that the radius r is negligible compared to the Compton wavelength of the charged particle one may neglect the particle mass, and then use dimensional analysis to see that the azimuthal current density must be proportional to r^{-3}. This implies a magnetic field proportional to r^{-2}, contributing a magnetic flux between shells of radii r and r' proportional to $ln(r'/r)$. Thus, as a test charge approaches the string, the apparent flux instead of remaining constant exhibits an anomalous dimension, familiar from the renormalization group treatment of electric couplings. However, still within the perturbative context, there is a big difference. The relevant beta function (to all orders in F, but lowest order in the gauge coupling α) vanishes for integer F. Since the AB coupling does not diverge, even though it does get strong enough to make perturbation theory suspect, it may be a more reliable indicator of the behavior in the large α regime than is the naively divergent α itself.

Dynamical strings of flux?

We may say with considerable assurance that in the strong coupling (large α) domain there will be large (order unity) fluxes present as vacuum fluctuations. Further, since at least in this abelian gauge theory flux is conserved, these fluxes in the vacuum plausibly could be excited to form observable moving strings of flux. However, if the net flux in such a string were an AB quantum (Here 'net flux' includes the accompanying vacuum-current induced flux), then low energy electrons would be insensitive to its presence, since if the interior is not penetrated an AB quantum is invisible. If net flux quanta other than zero were present in the vacuum fluctuations, the vacuum would exhibit 'spontaneous electric charge quantization', in the sense that a particle with a fraction of an electron charge would find its effective mass raised to a value on the scale where the coupling becomes strong.

The picture of a vacuum containing magnetic flux strings with diameter characterized by the strong coupling scale has been proposed before, by Nielsen and Olesen,[16] who used an intricate pattern of deduction to argue for the necessity of such a flux spaghetti in the nonabelian theory QCD. From the renormalization group point of view adopted here their argument seems quite natural. In QCD one may characterize flux in a gauge invariant way by obtaining the Wilson loop function, the trace of the gauge transformation associated with a particular loop in space. If that gauge transformation is a multiple of the unit operator, then the suitably normalized trace has possible values $1, e^{2\pi i/3}, e^{-2\pi i/3}$. Since gluons are insensitive to the presence of any such flux quantum, one may wonder if the full beta function might vanish at such values, leading to an enhanced likelihood of finding quantized fluxes, and therefore density enhancements in the complex plane near the above-mentioned values (for which the group invariant density actually vanishes).

If we now introduce quarks, which lie in the fundamental representation of $SU(3)$, then the beta function will vanish for Wilson loop trace 1. However, for each of the other two unit-matrix values, which would give a nontrivial Aharonov-Bohm effect on the quarks, their weaker and nonvanishing beta function will oppose the contribution from gluons, so that one expects the enhancements in concentration to be near but not at these values. Nevertheless, such a pattern would imply strong, locally correlated, color magnetic fields. These could well be a (or the!) critical factor generating color electric confinement.

Consistency of QED?

At this point let us pause and take stock. We have been treading familiar ground in the sense that it has long been known that couplings in perturbative field theories generally have anomalous dimensions which give rise to increasingly strong interactions as length scales get larger (QCD) or smaller (QED). If the coupling studied is an Aharonov-Bohm coupling, then at least in perturbation theory it is not divergent, but only approaches unit strength. This invites us to consider the AB coupling as a more reliable indicator of the true dynamics in the strong-coupling regime than the perturbatively divergent ordinary electric coupling. However, all of this assumes that QED, in particular, is a consistent theory. What is known about that?

Two different approaches give different answers. On the one hand, we know that a theory extrapolated to a new domain may fail either because it lacks essential physics, as Newton's mechanics fails for relativistic velocities, or because the theory is not consistent, with the inconsistency becoming dramatic in the new domain. There is strong circumstantial evidence that $\lambda\phi^4$ theory is consistent only for $\lambda = 0$. In perturbation theory there is a divergence as distance scales grow smaller, just as in QED. This suggests that QED may be viewed as a 'cousin' theory with the same genetic disease. However, the pathology in $\lambda\phi^4$ is such a borderline effect that it is easy to imagine a cure resulting from very slight changes. On the other side of this question, we have even stronger circumstantial evidence that QCD is consistent. Thus it becomes a question of which cousin theory is closer in its essence to QED, $\lambda\phi^4$ or QCD. If we choose the latter, then we need ask only the more limited but still quite challenging question, "What are the dynamics of QED on short distance scales?" For that, the well-behaved AB coupling is an appealing guide.

Let us explore the consequences of assuming that tubes of flux become dynamical degrees of freedom on the scale where the coupling is strong. Such a tube would have transverse dimensions in its rest frame, and also string tension, determined by the strong-coupling scale. A charged particle localized on this scale would receive arbitrarily large contributions to its effective mass from virtual flux strings of arbitrary velocity passing by. On the other hand, a spread-out particle wave function would be insensitive to these strings with their quantized flux. Thus the effect of this assumed vacuum structure would be to make sufficiently localized particles so massive that there would be negligible contributions from large virtual masses in loop diagrams for vacuum polarization.

Now we may consider whether there is a mechanism to generate the assumed flux tubes. Suppose a localized pair of electron and positron appear. If they overlap spatially, then they have negligible Coulomb energy. If further they have parallel magnetic moments then the energy is much lower than for antiparallel moments, so that the fluctuation should last longer. Furthermore, reinforcing fluctuations at neighboring locations are favored for the same reason. Thus correlated flux fluctuations corresponding to virtual flux strings seem inevitable. There is an extra subtlety in this argument. The notion of a magnetic moment is only simple in a nonrelativistic context, but that is immediately applicable here since it is being supposed that the electron becomes massive at the strong interaction scale. Thus the hypothesis that QED is consistent and that small diameter flux strings populate space leads to a picture of the vacuum which indeed hangs together, with the flux strings giving mass to the electrons and so halting the divergence of the ordinary electric coupling at small distances, and the massive, strongly-coupled electrons easily generating the flux strings.

Having speculated this far, let us go a little farther. For pure QED with one species of electron the scale factor for going from the electron Compton wavelength

λ_e to the strong coupling region is $e^{-3\pi/2\alpha} = 10^{-300}$. However, in the standard model, with $U(1)$ of QED replaced by $U(1)_R$, and three generations of quarks and leptons included, the number in the exponent is reduced by an order of magnitude, and comes to the vicinity of the value appropriate for the ratio of the Planck length to λ_e. Thus relatively minor modifications of the standard model could lead to flux strings appearing on the Planck scale. Such a circumstance would engender a temptation to identify the flux strings with the strings of string theory, which then could be treated as derived objects. If known light particles were derived from string theory, this could become the ultimate bootstrap!

Coda

We have arrived in the land of pure fantasy, but the fact that such a fantasy even could be conceived is a testament to the reach and scope of the AB effect, which at least perhaps might be not only the essence of gauge interactions but also the root of the whole structure of the Universe. This makes it quite fitting to close with a few personal remarks about the discoverers of the effect.

I met Yakir Aharonov a while ago, and by now have had a number of chances to experience his unique, adrenalin-raising approach to science. As with Niels Bohr's lucky horseshoe, it is not necessary to believe Yakir's ideas in order to benefit from them. Many others here can attest that even when disagreeing with him one finds he has exposed deep aspects of physics whose further study is bound to be fruitful. He comes closer than anyone I know to making the Socratic method a workable tool for learning about Nature. It is no surprise that this meeting in his honor should exhibit the same quality of excitement and discovery which we have learned to associate with Yakir.

I looked forward to this conference as my first opportunity to meet David Bohm, whom I had admired since college. I took a course on quantum mechanics in which the lecturing did not match my learning style very well, and his book was my salvation. For reading by oneself the high ratio of words to equations proved quite congenial, leaving me at the end feeling that I had grown up knowing quantum mechanics. That foundation has served me well ever since, and I am most grateful for it. Its author, by showing much more courage than I in probing and questioning the structure of quantum mechanics which he understood and explained so well, only increased my admiration for him as a person perpetually restless in the search for truth.

The paper of Aharonov and Bohm may have been the first scientific article I read on my own rather than for a class assignment. I remember being impressed by the striking simplicity of the argument but a little cautious because of the audacity of the language. When Furry and Ramsey[17] wrote a paper in response, the rumor I got from fellow students was that they had put Aharonov and Bohm in their place, demolishing the idea. Of course, when I read the paper it became clear that wasn't so. Instead they showed that the AB effect is necessary for the consistency of quantum mechanics, in particular for the complementarity between observation of wave interference and detection of particle trajectories. The whole episode was a wonderful introduction to science at the frontier, and shaped the work of many people. Not least significant is the fact that the AB phase was contained in a paper published ten years before, by Ehrenberg and Siday,[18] who seemed to take the effect as a matter of course and thus failed to focus on it the attention which it so richly deserves and has so richly repaid since 1959.

This work was supported in part by the National Science Foundation under Grant PHY 92-11367.

References

1 Y. Aharonov and D. Bohm, Phys. Rev. **115** (1959) 485.

2. A.S. Goldhaber, H.-N. Li, and R.R. Parwani, "Scaling of Aharonov-Bohm couplings and the dynamical vacuum in gauge theories", hep-th 9305007.

3. J. Anandan, Phys. Rev. D **15**, 1448 (1977), Internat. J. Theoret. Phys. **19** (1980) 537.

4. C.R. Hagen, Phys. Rev. D **41** (1990) 2015.

5. M. G. Alford, J. March-Russell and F. Wilczek, Nucl. Phys. **B328** (1989) 140.

6. C.R. Hagen, Phys. Rev. Lett. **64** (1990) 503, 2347; F. Vera and I.Schmidt, Phys. Rev. D **42** (1990) 35

7. Y. Avishai and Y.B. Band, Phys. Rev. Lett **66** (1991) 1761.

8. P. Górnicki, Ann. Phys. **202** (1990) 271.

9. R.R. Parwani and A.S. Goldhaber, Nucl. Phys. **B359** (1991) 483 (please note that Fig.2 in this paper is correct only for the induced density and not also for the current as claimed. The current decreases in the range $1/2 < F < 1$. The equations are correct. See also Ref. 10 below.); H.N. Li, D.A. Coker and A.S. Goldhaber, Phys. Rev. D **47** (1993) 694.

10. E.G. Flekkøy and J.M. Leinaas, Int. J. Mod. Phys. **A6** (1991) 5327.

11. E.M. Serbryanyi, Theor. Math. Phys. **64** (1985) 846.

12. V.S. Vanyashin and M.T. Terenteev, Sov. Phys. JETP **21** (1965) 375.

13. N.K. Nielsen, Am.J.Phys. **49** (1981) 1171.

14. R.J. Hughes, Nucl. Phys. **B186** (1981) 376.

15. Ph. Gerbert, Phys. Rev. D **40** (1989) 1346.

16. H.B. Nielsen and P. Olesen. Nucl. Phys. **B160** (1979) 380.

17. W.H. Furry and N.F. Ramsey, Phys. Rev. **118** (1960) 623.

18. W. Ehrenberg and R.W. Siday, Proc. Phys. Soc. London **B62** (1949) 8.

AHARONOV EFFECTS FOR TWO SLITS AND SEPARATED OSCILLATORY FIELDS INTERFERENCES

Norman F. Ramsey
Lyman Physics Laboratory
Harvard University
Cambridge, MA 02138, USA

ABSTRACT

The implications of complementarity on two path interferences and separated oscillatory field resonances are discussed. Furry and Ramsey have shown that an apparatus to determine the electron path introduces uncertainties in the scaler and vector potentials that disturb the phase of the electron wave function so much through the Aharonov-Bohm effects that the interference fringes disappear. A similar result is derived for the neutron, but with the phase uncertainties coming from the magnetic moment's motion through an electric field discussed by Anandan, Aharonov and Casher. The separated oscillatory field resonance method can be interpreted as an interference between two different paths in spin space. The same analysis as for the neutron two path interferences shows that the separated oscillatory field resonance disappears when the orientation of the neutron spin is observed between the two oscillatory field regions. An interesting difference between the separated paths and separated oscillatory fields experiments is that the latter may be interpreted classically. An equal superposition of the two orientation states along one axis corresponds to an eigenstate relative to an orthogonal axis so the separated oscillatory field resonances can be interpreted classically whereas this is not possible with the two path interferences.

1. Introduction

It is a pleasure to speak at this conference honoring Y. Aharonov, whose stimulating papers have added so much to our understanding of quantum mechanics but I deeply miss David Bohm.

I remember well the waves of surprise and disbelief that circulated throughout much of the physics community on the 1959 publication of the Aharonov-Bohm (AB) paper[1] that pointed out the possibility of observable effects of electromagnetic potentials on charged particles unexposed to electric or magnetic fields. Wendell Furry and I at that time were astonished but willing to try to understand the effects from different points of view. As a result we published one of the first papers supporting the AB analysis[2]. We pointed out that the (AB) effects for scalar and vector potentials were essential to preserve the consistency of quantum mechanics and the principle of complementarity. We showed that, without these effects of the scalar and vector electromagnetic potentials, it would be possible to observe two slit interference patterns with charged particles while at the same time detecting through which slit the particle went. Such an observation is inconsistent with the principle of complementarity applied to a two slits interference experiment[2]. We showed that the AB effects would make the interference pattern disappear if the path detection sensitivity were sufficient to determine through which slit the charged particle went.

Since our early paper convinced many scientists of the validity of the AB observations, the organizers of this conference urged me to review that paper here. However, I was reluctant to repeat a 32 year old paper in a field in which I have done no recent work. But, I then realized I could also analyze two different problems from a similar point of view, so I agreed both to review our old paper and to discuss the new subjects, even though the three different reports produce a cumbersome collective title. The two new analyses depend on the phase shifts of a neutral particle with a magnetic moment moving through an electric field as discussed by Anandan[3], Aharonov[4], and Casher[4] (AAC). The first of the three reports reviews our old work under the title Complementarity and Two Paths Electron Interferences . The second is Complementarity and Two Paths Neutron Interferences and the third is Complementarity and Separated Oscillatory Fields Resonances.

2. Complementarity and Two Paths Electron Interferences

The AB paper[1] considered the effects of both the scalar and the vector electromagnetic potentials so Furry[2] and I did likewise. In the case of the scalar potential we considered the idealized apparatus shown in Figure 1 to see if it could be used to detect through which slit the electron passed while still observing the interference pattern. The detection of the slit traversed by the electron is made by determining which way the test body of charge q is accelerated before the electron emerges from the pipe. The test body is between two condenser plates separated by a distance l as in Figure 1. It is held fixed half way between them ($x = l / 2$) until the waves are inside the tube, and is brought back to this position before the waves emerge; thus it produces no field between the pipes at any time when the field could act on

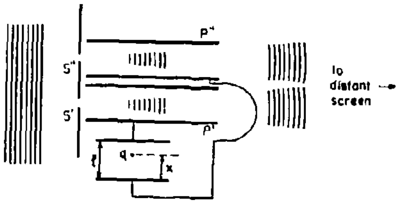


Figure 1. Electrostatic effects.

the particle. The test body is free to move during a time interval T when the waves are certainly inside the pipes, and by determining the direction in which the test body is accelerated during the time T we can find out which tube contains the particle.

The potential difference produced by the presence of the electron in one tube or the other is $V_1 = \pm e / (2C)$, where C is the total capacity of the condenser and attached pipes. The magnitude of the field strength is thus[2]

$$|\mathbf{E}| = e / (2lC). \qquad (1)$$

The force on the test body is q **E**. If its direction is to be determined, it must produce a change of the momentum of the test body that is larger than the uncertainty of that momentum Δp. To be relatively certain of the direction we take the imparted momentum to

$$q\,|\mathbf{E}|\,T > 2\,\Delta p. \qquad (2)$$

Displacement of a charge q from a central position at $x = l / 2$ produces a potential difference[2]

$$V = (q / C)(x - l / 2) / l \qquad (3)$$

and the uncertainty of the potential difference is

$$\Delta V = (q / lC)\Delta x \qquad (4)$$

Substituting Eq. (1) into (2) and multiplying by Eq. (4), we have

$$q\,e\,T\,\Delta V / (2lC) > 2(q / lC)\,\Delta p\,\Delta x \geq 2(q / lC)\,h / (4\pi) \qquad (5)$$

Therefore,

$$e\,T\,\Delta V > 2\,h / (2\pi). \qquad (6)$$

By AB[1], if alternative electron paths involve the electron being in electrostatically shielded regions with a potential difference V for a time T the wave functions will develop a difference of phase of $e V T / (h / 2\pi)$.

Therefore the uncertainty in the phase differences between the two paths caused the test body is

$$\Delta\phi = e T \Delta V / (h / 2\pi) > 2 \qquad (7)$$

A phase shift uncertainty of 2 radians will obliterate the fringes, so the AB effect of the electrostatic potential assures the consistency of quantum mechanics by making it impossible to obtain interference fringes when the electron path is known.

When I first reported on our work at scientific meetings in 1959, I introduced into the scientific literature a Charles Addams cartoon which has since been used extensively. This cartoon shown in Figure 2 is a great illustration of a fundamental difference between classical and quantum mechanics. In a classical world the cartoon is a joke since a classical object can not possibly pass through two separated regions at the same time. On the other hand in a quantum world the wave function of an electron can simultaneously experience the potentials at two separate regions of space.

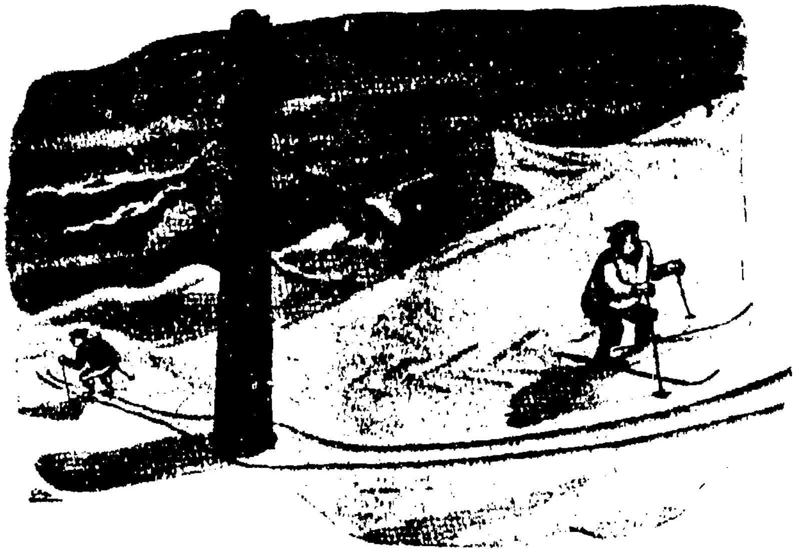


Figure 2. Charles Addams cartoon.

With the AB vector potential effect, the analysis is similar but a coil and an infinitely long infinitely permeable rod R are used for the path detection as shown in Figure 3. The coils and plates are assumed to have no resistance. With these assumptions there is no stray flux outside the rod and hence no field in the regions traversed by the electrons. Furthermore the current induced in the search coil is then just that required to prevent any change in the flux Φ. Passage of the particle through either slit and on to the screen is tantamount to flow of the charge e through one half turn since a trip out and symmetrically back on the other side would be equivalent to a full turn. Therefore, with N coil turns, the charge delivered to C is[2]

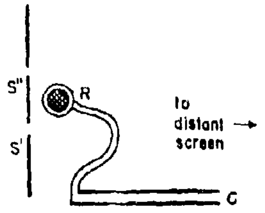


Figure 3. Magnetic effects.

$$Q = \pm e \,/\, 2N \tag{8}$$

The characteristic time of the circuit[2]

$$T = (L\,C)^{1/2}/c \tag{9}$$

is very long compared with the time of passage of the wave packet through the apparatus. Thus we can have the advantage, as compared to the scalar potential analysis, of ample time for the determination of the sign of Q.

The circuit has two canonically conjugate variables, the charge Q and the flux linkage N Φ, which appear in the Hamiltonian for the equivalent harmonic oscillator,

$$H = Q^2 \,/\, 2\,C + (N\,\Phi)^2 / 2\,L, \tag{10}$$

and satisfy the uncertainty relation

$$\Delta Q\; N\; \Delta\Phi > hc \,/\, 4\,\pi. \tag{11}$$

If we are to determine the sign of Q reliably, we must have

$$|Q\,| > 2\,\Delta Q\,. \tag{12}$$

From this and Eqs. (8) and (11) we obtain

$$e\;\Delta\Phi \;> 2\;hc \,/\, 2\;\pi, \tag{13}$$

AB point out that there is a resultant phase difference

$$\phi \;=\; (\,2\pi\,e \,/\, h\,c\,)\;\Phi \tag{14}$$

between the waves that have passed R on one side or the other. Consequently from Eq. (13) the spread in ϕ is given by

$$\Delta\,\phi = (\,2\;\pi\,e \,/\, h\,c\,)\;\Delta\Phi > 2 \tag{15}$$

3. Complementarity and Two Paths Neutron Interferences

The above AB analysis applies only to charged particles. However two slit interference patterns for many years have been obtained with neutrons and more recently with neutral atoms. One can attempt to to detect through which slit a polarized neutron went while observing the interference pattern. Various methods can be chosen to detect the neutron path and for each there is a corresponding uncertainty relation that destroys the interference pattern. For example, an apparatus similar to that shown in Figure 3 could be used but with neutrons polarized perpendicular to the paper and with neutrons on one possible path having to pass through the infinitely permeable rod. The sign of the charge Q delivered could then be observed as in the vector potential case discussed above. However, as in that discussion the uncertainty in the magnetic field destroys the interference pattern when the apparatus is sufficiently sensitive to determine the path.

A different method of path detection uses the fact that a magnetic dipole of strength μ_M moving with velocity v appears in a stationary reference frame to have an electric dipole moment μ_E given by

$$\mu_E = (v / c) \times \mu_M \, , \qquad (16)$$

so the passage of the neutron through a condenser could be detected by measuring the induced potential. The same Figure 1 with a different interpretation can be used to describe the proposed experiment. Instead of the four dark horizontal lines being interpreted as two pipes, they now represent plates of two parallel plate condensers with the inner two plates connected together. The neutrons are polarized perpendicular to the paper so the sign of the potential induced by a passing neutron depends on which slit is traversed. From Eq. (3) applied to each pole of an electric dipole, it can be seen that the potential V_1 induced during the passage of an electric dipole through one condenser or the other is

$$V_1 = \pm \mu_E / (C d) \qquad (17)$$

where d is the separation of the plates in the condenser. The magnitude of the field E on the test charge is then

$$| \mathrm{E} | = V_1 / l = \mu_E / (Cdl) = v \; \mu_M / cCdl \qquad (18)$$

To be relatively certain of the direction q moves as in the AB electrostatic discussion, we must have

$$2 \Delta p < \mathrm{p} = \mathrm{q} \, | \mathrm{E} | \, \mathrm{T} = q \, v \, \mu_M \; T / c C d l \approx \mathrm{q} \; \mu_M \; L / c C d l \qquad (19)$$

where L = Tv is the length of the condenser.

But the detection mechanism in Figure 1, by Eqs. (4) and (19) will have an uncertainty in voltage of

$$\Delta V = q\, \Delta x / l C > q h / 4 \pi l C \Delta p > h d c / 2 \pi \mu_M L \quad (20)$$

But by AAC[3,4] and Cimmino et al[5] the phase difference between the two sides is

$$\phi_{AAC} = (2 \pi / h) \int p \cdot dr = 4\pi \mu_M \Lambda / c (h / 2 \pi)$$

$$= 2 \mu_M L V / d c (h / 2 \pi) \quad (21)$$

where Λ is the lineal charge density which by the Gauss theorem is related to the voltage across the condenser by

$$\Lambda = 2 V L / 4 \pi d. \quad (22)$$

The uncertainty in the phase by Eq. (20) then is

$$\Delta\phi_{AAC} = 2 \mu_M L \Delta V / d c (h / 2 \pi) > 2. \quad (23)$$

So the interference disappears just as the path of the neutron is detected.

4. Complementarity and Separated Oscillatory Fields Resonances

For the resonance methods of separated or successive oscillatory fields[6,7], the transition probabilities for a two level system can be exactly calculated. Although the resulting formulae are useful in determining spectral line shapes, they obscure the origins of the observed sharp peaks as coming from an interference between two possible paths in spin space. However, this origin can be clarified by deriving the transition formula in an alternative, but equivalent form. For simplicity the same notation will be used as in the original papers[6,7] and consideration will will be restricted to the special limiting case where the durations τ of the two pulses are negligibly small except when multiplying the transition inducing amplitude b, which is assumed so large that $b\tau$ is finite.

With these restrictions, the exact expressions[6,7] for the probability amplitudes after the interaction of duratic n τ in terms of those at times t_1 just before the interaction simplify to

$$C_p(t_1 + \tau) = \cos(b\tau) C_p(t_1) - i \sin(b\tau) \exp(i\omega t_1) C_q(t_1) \quad (24)$$

$$C_q(t_1 + \tau) = -i \sin(b\tau) \exp(-i\omega t_1) C_p(t_1) + \cos(b\tau) C_q(t_1)$$

However , following a finite period T with b = 0, the probability amplitudes are related as follows:

$$C_p(t_1 + T) = [\exp(-i 2 \pi W_p T / h)] C_p(t_1) \quad (25)$$

$$C_q(t_1 + T) = [\exp(-i 2 \pi W_q T / h)] C_q(t_1).$$

By successively applying these relations it is easily seen for $C_p(0) = 1$ and $C_q(0) = 0$ that

$$C_q(T + 2\tau) = -i \sin(b\tau) \cos(b\tau) [\exp - i (\omega + 2 \pi W_p / h) T$$
$$+ \exp - i (2 \pi W_q / h) T] \quad (26)$$

The modulus squared of Equation (26) gives for the transition probability

$$| C_q(T + 2 \tau) |^2 = 4 \sin^2(b\tau) \cos^2(b\tau) \cos^2 (\omega - \omega_0)T/2 \qquad (27)$$

which is in agreement with the usual expression[4,5] when the above restrictions are applied.

In Eq. (26) the first term corresponds to the probability amplitude for passing through the intermediate region in the original state p followed by a transition in the second oscillatory field. The extra factor $exp - i (\omega) T$ arises from the phase of the oscillatory field at the time T when the transition occurs. The second term corresponds to the probability amplitude of a transition to state q in the first transition region with passage through the intermediate region in state q. From the form of Eq. (26), it is apparent that the factor $\cos^2 (\omega - \omega_0)$ comes from the cross terms between the probability amplitudes for the two possible spin orientation paths between the two oscillatory field regions.

In the case of the separated oscillatory field method, the analogue to determining through which slit the particle passes is determining the spin orientation state of the particle during the interval between the two coherent pulses. In the case of neutrons this might be done by allowing the beam to pass between two plates of a condenser and determining the orientation state from the sign of the induced potential as in the previous discussion. The analysis is the same as for the two slit case and the sharp resonances disappear through the AAC effect just as the sensitivity becomes sufficient to detect the orientation, as required by complementarity. Englert, Walther and Scully[8] have recently and independently made analogous observations using a micromaser with two field optical fringes.

Despite the similarities, there are fundamental differences between the two slits and the separated oscillatory fields experiments with neutrons. The orientation state of the neutron is determined by a vector in three dimensions and an equal superposition of the m = + 1/2 and -1/ states corresponds to an orientation eigenstate along an axis perpendicular to the original axis. As a result the sharp resonance peaks can be interpreted classically as the spin being flipped π/2 radians in the first oscillatory field and being allowed to precess before the next one. If the precession and oscillator frequencies are the same, the second osciallating field will do the same thing as the first, producing a maximum reorientation. If on the other hand the frequencies are slightly different so that the neutron spin precesses an extra π radians, the spin will be flipped back to its original position corresponding to a minimum transition probability, thus providing narrow resonance even with a classical interpretation. On the other hand, such a classical interpretation of two slit interferences is not possible since there is no reasonable classical interpretation for the probability amplitude corresponding to the superposition of two different paths in space.

References

1. Y. Aharonov and D. Bohm, *Phys. Rev.* **115** (1959) 485. Following references 1 and other papers in this field, Gaussian units are used her .
2. W. H. Furry and N. F. Ramsey, *Phys. Rev.* **118** (1960) 623.
3. J. Anandan, *Phys. Rev. Letters* **48** (1982) 1660.
4. Y. Aharonov and A. Casher, *Phys. Rev. Letters* **53** (1984) 319.
5. A. Cimmino, G. I. Opat, A. G. Klein, H. Kaiser, S. A. Werner, M. Arif and R. Clothier, *Phys. Rev. Letters* **61** (1989) 380.
6. N. F. Ramsey, *Phys. Rev.* **78** (1950) 695.
7. N. F. Ramsey, *Molecular Beams* (Oxford University Press, 1956, 1990), pps. 127-128.
8. B. G. Englert, H. Walther and M. O. Scully, *Appl. Phys.* **B 54** (1992) 366.

ATOM INTERFEROMETERS

David E. Pritchard
Department of Physics and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

1. Introduction

It is a pleasure to speak at Yakir's birthday celebration, especially on the subject of matter wave interferometers which his work has been intimately related to for so many years. And the fact that we will be able to give the first report on experiments in which atoms are sent on both sides of a metal foil and then recombined adds even more enjoyment. But let me begin this talk from the atomic physics point of view.

The field of atomic, molecular, and optical physics has been moving with increasing velocity in recent years, and no subfield in this area is currently developing faster than atom optics and atom interferometers (both of which we have recently reviewed.[1,2]) About a dozen experiments that demonstrate atom wave interference or atom interferometers have been performed during the last five years;[3-11] results which parallel demonstrations of optical devices that spanned most of the nineteenth century. Even without further developments in atom optics, there now exist enough useful elements to make a variety of atom interference devices and interferometers. The burst of activity in this area in 1991 was reported in most of the widely circulated scientific maga ines,[12] as well as in recent review articles.[2,13,14] Since these reviews, two new groups have performed related demonstration experiments.[10,15] More importantly, measurements are now being made using these devices. At this rate, we anticipate that many applications of atom interferometers to probl ms of scientific and technical importance analogous to those of the last hundred years in optics will be made in the remainder of the 1990's. (This is not to say that we're smarter, just that we have the theoretical understanding to chart a surer course, and much commercially available technology with which to pursue it rapidly.) Since our judgement is that we have now entered a period in which the most important advances involving atom interferometers will be new applications rather than new interferometers, the remainder of this presentation will concentrate on the four areas in which atom interferometers appear likely to have significant scientific and technical applications. Unfortunately for those who have read our recent review,[2] there is little new to report in the couple of months since that was written.

2. Atom Interferometer Applications

2.1 Atomic and Molecular Properties

We are pleased to report at this conference the first measurements made with a separated beam atom interferometer. The key point here is that separated beam atom interferometers present the opportunity to subject part of the atom wave to an interaction which causes a phase shift and then to measure this phase shift by interference with the unshifted part of the atom wave in the same state. An obvious application is the precision measurement of the (ground state) polarizability of the atoms (or molecules) in the interferometer, by subjecting one part of the atom wave to a uniform electric field. We discuss such an experiment currently giving nice results in our laboratory. We note that this is intrinsically a higher precision approach than measuring the deflection in a field gradient, whether this is measured by conventional[16] or interference techniques.[17] We also note that *polarizability differences* between two states of an atom have long been measured using optical resonance techniques, and can also be measured in interferometers of the Chebotayev type even though the two legs of the interferometer are not spatially resolved.[15,18]

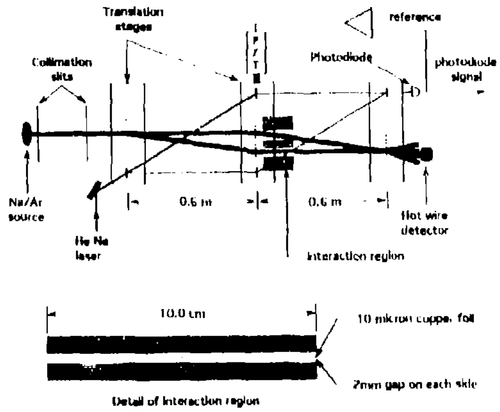


Figure 1: A schematic, not to scale, of our atom interferometer. The 10 μm copper foil is between the two arms of the interferometer (thick lines are atom beams). The optical interferometer (thin lines are laser beams) measures the relative position of the 200 nm period atom gratings (which are indicated by vertical dashed lines).

Our atom interferometer has been described in,[3] and is depicted in Fig. 1. It uses three equally spaced transmission gratings, a standard interferometer design, with about 2/3 of a meter spacing between the gratings. This configuration produces a robust white fringe.[19] We now use three 0.2 micron period nanofabricated[20] diffraction gratings which separate the centers of the interfering beams by 55 μ. We collimated the sodium atom beam with 20 μ slits so the edges of the two interfering beams do not overlap. The atoms had a deBroglie wavelength of 16pm. The FWHM

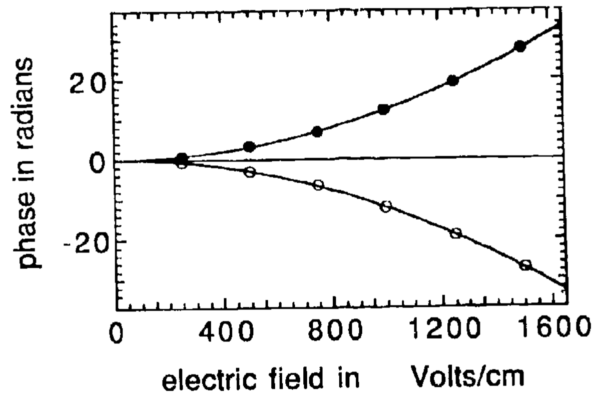


Figure 2: Interference pattern from 40 seconds of data (~ 1 second per point). The contrast is 25% and the phase uncertainty is 17 milliradians. Detector noise background of 200 counts per second has been subtracted.

of the velocity distribution of the beam was 11%, which determined the longitudinal coherence length (1.6 Å). The fringe amplitude was 820 cps, which would allow us to determine the phase to 15 milliradians in 1 minute (Fig. 2).

An interaction region consisting of a stretched metal foil positioned symmetrically between two side electrodes, each spaced 2 mm from this septum, was inserted in the interferometer so that the atom wave in the two sides of the interferometer passes on opposite sides of the foil. The septum was 10 cm long and 10 microns thick, but the shadow it cast on the detector was typically 30 μ wide due to slight deviations of the stretched foil from perfect flatness. Because we have a conducting physical barrier between the separated beams, we can apply different, but uniform, electric and magnetic fields to the portions of the atom wave on each side of the interferometer.

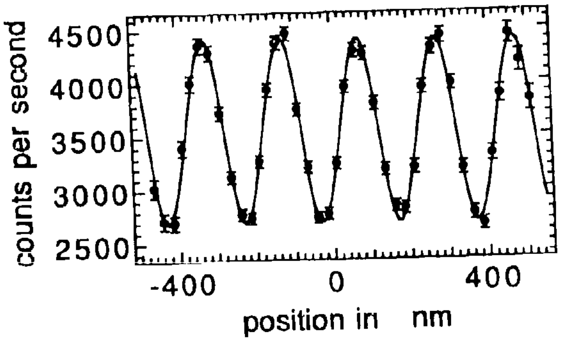


Figure 3: Stark phase shifts for voltages applied to the right (open circles) and the left (filled circles) side of the interaction region. Phase shift per applied electric field squared in (volt/cm)^2 is $1.220(7)\times10^{-5}$ for the left side and $1.224(7)\times10^{-5}$ for the right side. This measurement determines the dc polarizability of sodium with 0.4% statistical uncertainty.

If an electric field is put on one side of the interaction region the DC Stark shift of the atom wave on that side will change the phase of the interference pattern. The Stark shift is

$$V_s = -\alpha \varepsilon^2/2,$$

where α is the electric polarizability and ε is the applied electric field. The Stark shift acts as a slight depression in the potential energy, V, as the atom wave passes through the electric field. This increases its spatial frequency (since the deBroglie wave number is $k = [\,2m(E - V)]^{1/2}/\hbar$ and E is conserved), resulting in an increased phase accumulation relative to the wave that passes on the side of the septum with no field. Since V_s is eight orders of magnitude smaller than E, the square root can be expanded with the result that the differential phase shift is $\phi = \alpha l\, \varepsilon^2/2v\hbar = V_s t/\hbar$ where l is the length of the interaction region, v is the velocity of the atoms, and t is the transit time. We found that the measured phase shift was quadratic with the applied field within error, as expected, allowing us to determine the polarizability of the ground state. We found that putting the field on opposite sides of the septum gave the same absolute value of the phase shift, giving a statistical error of .4% in ~ 20 minutes. We are currently investigating several systematic errors (the largest due to variation of the phase shift with septum position) which currently limit our determination of an absolute value of the polarizability to ~ 1%. Figure 3 shows the phase shifts vs. applied electric field.

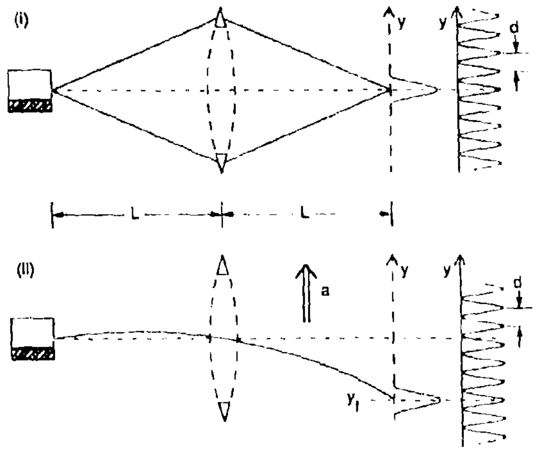


Figure 4: Contrast revivals from constructive rephasing of the independent interference patterns of the 8 different magnetic sub-states of sodium. These patterns are dephased by a current flowing down the septum which alters the magnitude of the uniform magnetic field on the two sides.

We have also observed the periodic rephasing of the independent interference patterns of the different Zeeman substates of the ground state as a differential magnetic field is applied to opposite sides of the septum. To observe this, we first applied a uniform magnetic field along the beam axis to determine the quantization direction. By running a current down the metal septum, perpendicular to the plane of the interferometer, we increased the field magnitude on one side of the interaction region and decreased it on the other. This gave a differential Zeeman energy, and therefore phase, for the two paths around the foil. The phase shift is proportional to the current passing through the septum and the projection of the magnetic moment along the quantization axis. Since our beam is unpolarized, the observed interference pattern is a sum of interference patterns for each of the eight sodium ground states. Since the g-factors of the F = 1 and F = 2 hyperfine levels have equal magnitude (but opposite sign), there are only three different magnitudes of projected magnetic moment. At low fields these are proportional to 0, 1/2, and 1 times a Bohr magneton. Consequently the independent interference patterns periodically rephase constructively to produce a high contrast interference pattern with the same phase as the pattern at zero field. This is shown in Figure 4. The first revival of contrast is the point where the phase shifts are 4π for the $|m_F| = 2$ states, 2π for the $|m_F| = 1$ states, and 0 for the $m_F = 0$ states. In this experiment, therefore, the informative variable is the contrast (not phase) vs. magnetic field.

The contrast versus differential magnetic field has the same shape as the amplitude versus position for a five slit diffraction grating whose central three slits are twice as wide as the extremal slits. (To make this analogy more precise, we would have to illuminate the grating with light of the appropriate spectral width.) Fig. 4 also contains a fit to the data which correctly models the effects of our finite velocity distribution and misalignment of the uniform magnetic field that determines the quantization axis. Not only the relative positions of the contrast maxima, but also their width and the degradation of contrast of subsequent rephasings due to the finite coherence length is well accounted for by the model. The real significance of this rephasing experiment is that (since the value of the Bohr magneton is accurately known) one of the fit parameters is the average velocity of the atoms that successfully make it through the interaction regin and contribute to the interference pattern. This can be exploited to eliminate systematic effects arising from processes which cause this final average velocity to differ from the average velocity of the atoms in the beam upstream of the interferometer.

For large currents down the foil, the average over the velocity distribution of the atom beam reduces the contrast in the interference pattern of all atoms except those in the two $m_F = 0$ states, which experience no Zeeman phase shift. This will result in a contrast one-fourth of that observed for no current. At this point, any small phase shifts observed from additional interactions would be those of only the $m_F = 0$ states. By applying a large Stark phase shift to all of the substates, the contrast of these $m_F=0$ states could be reduced to nearly zero while another polarization state was shifted back into coherence with itself. This would allow experiments to be performed on a polarized beam without the difficulty of optical pumping, (but without the gain in intensity which such optical pumping should bring).

What happens if the atom wave on one side of the septum passes through a gas not present on the other side? From the perspective of wave optics, the passage of a wave through a medium is described in terms of an index of refraction whose real part is proportional to the phase shift of the wave and whose imaginary part is proportional to the absorption of the wave. For an atom wave passing through a gaseous medium, absorption will be proportional to the well understood total scattering cross section, which is determined by the imaginary part of the scattering amplitude at zero angle. The phase shift will be proportional to the real part of the scattering amplitude at zero angle. Taken together, the absorption and phase shift therefore determine the phase and amplitude of the scattering amplitude. In low energy collisions this means that both the magnitude and sign of the scattering length can be determined, an important advance since knowledge of the sign, hitherto not measurable, is critical to predicting low temperature collective behavior.

2.2 Inertial Effects

Atom interferometers are sensitive to inertial effects because the atoms travel freely (if field gradients are sufficiently small), whatever the acceleration of the apparatus. The difference in position of the interference pattern when observed in an accelerating vs. an inertially stable apparatus can be observed interferometrically, giving a precise measure of the non-inertial behavior of the apparatus. To make these ideas quantitative, first imagine atoms with velocity v passing through a matter wave lens with focal length L/2 as shown in Fig. 5; if the apparatus accelerates upwards at a, the central atom ray appears to follow the curved path shown, and the position of the image of the source will have a vertical displacement,

$$y_f = v_{yo}(2t) - 1/2\ a\ (2t)^2 = -at^2 = -\ a(L/v)^2\ ,$$

where t is the flight time for distance L and v_{yo} is the initial y velocity necessary to pass through the center of the lens. If the lens is converted into a separated Fresnel biprism by blocking off its central dashed portion, the Airy diffraction pattern of the lens will be converted into an extended interference pattern and the shift in position of the central fringe can be measured as a phase shift,

$$\phi = 2\pi y_f/d = -2\pi a t^2/d,$$

where d is the fringe spacing. The above expression also applies to a three grating interferometer with grating period (or lattice spacing) d.

The equivalence principle dictates that the response of an apparatus with acceleration g upwards must be the same as a stationary apparatus in a downward gravitational field with strength g. Thus the phase shift in a gravitational field should be

$$\phi_g = -\frac{2\pi g}{d}\left[\frac{L}{v}\right]^2$$

This result (with appropriate trigonometric modifications for finite opening angle) has been checked using neutrons;[21] a small discrepancy exists. The application of more complex interferometer configurations to the determination of the gravitational gradient has also been discussed.[22]

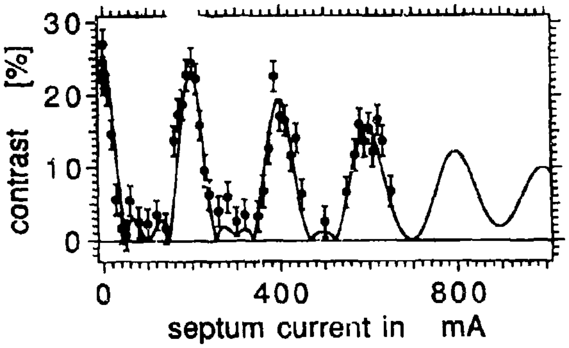


Figure 5: With no acceleration (i), the atom image (dashed axis) from the lens (dashed) and the diffraction pattern (solid axis) from the Fresnel biprism (solid lines at edges of lens) line up at y = 0. When the apparatus accelerates upward (ii), image and diffraction pattern are both displaced by y_f. (see Eq. 3).

If the apparatus is rotating with angular velocity $\vec{\Omega}$, the atoms experience a Coriolis acceleration $a_C = -2\,\vec{\Omega}\times\vec{v}$. For the interferometer discussed above, the phase shift due to this rotation may be calculated by substituting this acceleration into Eq. 4 with the result,

$$\phi_C = \frac{4\pi L^2}{dv}\,\Omega\,,$$

assuming small opening angles in the interferometer. This result has been verified for both neutron[22] and atom interferometers.[7]

Although this equation expresses the phase shift in terms of the experimentally specified parameters, it is customary to express the phase shift due to rotation in terms of the enclosed area, A, by the atoms. For grating type interferometers this is determined by the diffraction angle, $\beta = \lambda_{dB}/d$, yielding $\phi_C = 2m/\hbar\ \vec{\Omega}\cdot\vec{A}$, the familiar Sagnac phase for matter wave interferometers.[24,25]

Atom interferometers cannot measure any new inertial effects intrinsic to atoms, so the real question is one of technical performance. For rotation sensing, the greater phase shift of matter wave interferometers relative to light interferometers of the same configuration (by the factor $mc^2/\hbar\omega_{light} = 10^{10}$) suggests that improved atom optics technology (especially a non-diffractive beamsplitter) should enable atom interferometers to attain better precision than laser gyros. For measurement of the local gravitational constant (or for accelerometers) the demonstration of sensitivity of 3×10^{-8} in the first slow atom interferometer by the Stanford group[26] is very encouraging, especially if further experiments verify the projected freedom from systematic error.

2.3 Fundamental Measurements

The inherent precision available with interferometry makes atom interferometers ideal instruments with which to make fundamental "null" tests (e.g. of the charge of a neutral atom). Sensitivity to phase (as opposed to energy times time) will allow atom interferometers to probe physical processes that generate phase shifts such as Berry's (and other) topological phases (cf. a recent related proposal,[27] also discussed at this conference), the passage of atoms through a waveguide, or the phase shift which accompanies surface bounces. In general it has not previously been possible to observe these phase-generating effects.

A recent proposal by Anandan[28] and Aharonov and Casher[29] combines two of these ideas: it is a topological phase which tests a fundamental tenet of quantum mechanics — that a phase shift can occur in the absence of any classical force. A study of this effect using neutron interferometers is presented in these proceedings,[30] so we need not dwell on its desirability here. The advantages of using atoms are the greater magnetic moment (partially offset by the large Stark shift which limits the practical size of the fields which can be applied) and the greater intensity. Together these should greatly reduce the statistical error and should also allow us to study, for the first time, the predicted dependence on the dipole orientation.

Another important measurement is the precise determination of the momentum of a photon; an experiment underway at Stanford has been described.[31]

Before getting too carried away with the possibilities of new fundamental measurements, we should note that many fundamental experiments in matter wave optics and matter wave interferometry have already been carried out using neutron interferometers. A recent review of this work[32] serves as both a source of inspiration and a standard of comparison in this field.

2.4 Direct Write Atom Holography

Looked at from another perspective, our three grating interferometer is a holographic apparatus that produces a real image in the plane of the third grating. By changing the geometry (e.g. using the two first order beams from the first grating and the second order beams at the second), this image can be made to differ from the gratings used upstream (in this example it would be a grating with half the period of the others). If the middle grating were replaced by a calculated hologram (this would be easy since the electron beam writer which writes the grating[20] is computer controlled), the resulting image could be quite arbitrary. Recently it has been shown[33] that an atom image like the one just described can be written on a substrate with resolution better than 3000 Å, so the possibility of writing patterns of a particular type of atom on a surface already exists. If some way were found to develop this image (if it were written in silver, regular photographic techniques might be applicable) it would be a directly written atom structure.

3. Summary

The future of atoms interferometers looks bright: atom beams are inexpensive and intense relative to neutron beams from reactors, several techniques have now been demonstrated to make interferometers for them, and the atoms which may be used in them come with a wide range of parameters such as polarizability, mass, and magnetic moment. One can even imagine applications for molecular interferometers. This assures the applicability of these instruments to a wide range of measurements of both fundamental and practical interest. Hence atom interferometers may now be regarded as devices to think up experiments for. Ultimately they should become sufficiently robust and simple that they can be regarded as instruments, to be applied technologically or used in other experiments.

Our recent work on atom interferometers and atom optics is supported by the Army Research Office contracts DAAL03-89-K-0082, and ASSERT 29970-PH-AAS, the Office of Naval Research contract N00014-89-J-1207, and the Joint Services Electronics Program contract DAAL03-89-C-0001. I am very grateful for all of this support, for the heroic work done by my many graduate students, for enjoyable and helpful discussions on the subject of this paper with C. Ekstrom and J. Schmiedmeyer and for help preparing this manuscript from M. Chapman and T. Hammond.

References

1. D.E. Pritchard, *Atom Optics*, ed. by J.C. Zorn, R.R. Lewis, ICAP 12 (American Institute of Physics, Ann Arbor 1991) pp. 165-174.

2. D.E. Pritchard, *Atom Interferometers*, in Proceedings of the 13th International Conference on Atomic Physics, Munich, Germany, August 3-7, 1992, ed. T.W. Hansch and H. Walther.

3. D.W. Keith, C.R. Ekstrom, Q.A. Turchette, and D.E. Pritchard, Phys. Rev. Lett. **66**, 2693 (1991).

4. D.W. Keith, M.L. Schattenburg, Henry I. Smith and D.E. Pritchard, Phys. Rev. Lett. **61**, 1580 (1988).

5. P.L. Gould, G.A. Ruff, and D.E. Pritchard, Phys. Rev. Lett. **56**, 827 (1986).

6. O. Carnal, J. Mlynek, Phys. Rev. Lett. **66**, 2689 (1991).

7. F. Riehle, Th. Kisters, A. Witte, J. Helmcke, J. Borde, Phys. Rev. Lett. **67**, 177 (1991).

8. M. Kasevich, S. Chu, Phys. Rev. Lett. **67**, 181 (1991).

9. J. Robert et. al., Eur. Phys. Lett. **16**, 29 (1991). 10. F. Shimizu, K Shimizu and H. Takuma, Phys. Rev. A**46**, 46 (1992).

11. U. Sterr, K. Sengstock, J.H. Muller, D. Bettermann and w. Ertmer, Appl. Phys. B**54**, 341 (1992).

12. F. Flam, Science, 921 (1991); B. Levy, Physics Today, 17 (1991).

13. Ch. Miniatura, J. Robert, S. LeBoiteux, J. Reinhardt and J. Baudon, Appl. Phys. B, special issue

"Optics and Interferometry with Atoms".

14. Y. Aharonov and A. Stern, Phys. Rev. Lett. **69**, 3593, (1992).

15. U. Sterr, K. Sengstock, J.H. Muller, D. Bettermann and W. Ertmer, Appl. Phys. B **54**, 341 (1992).

16. R. Molof, H. Schwartz, T. Miller and B. Bederson, Phys. Rev. A. **10**, 1131 (1974).

17. F. Shimizu, K. Shimizu, and H. Takuma, J. Appl. Phys. **31**, L436 (1992).

18. F. Riehle, A. Witte, Th. Kisters, and J. Helmcke, Appl. Phys. B **54**, 383 (1992).

19. B.J. Chang, R. Alverness, and E.N. Leith, Appl. Optics **14**, 1597 (1975).

20. C.R. Ekstrom, D.W. Keith, and D.E. Pritchard, Appl. Phys. B **54**, 369 (1992).

21. R. Collela, A.W. Overhauser, and S.A. Werner, Phys. Rev. Lett. **34**, 1472 (1975).

22. J.F. Clauser, Physica B **151**, 262 (1988).

23. S.A. Werner, R. Collela, A.W. Overhauser, and C.F. Eagen, Phys. Rev. Lett. **35**, 1053 (1975).

24. J. Anandan, Phys. Rev. D **15**, 1448 (1977).

25. L.E. Stodolsky, Gen. Relativ. and Gravitation **11**, 391 (1979).

26. M. Kasevich, S. Chu, App. Phys. B **54**, 321 (1992).

27. Ady Stern, Phys. Rev. Lett. **68**, 1022 (1992).

28. J. Anandan, Phys. Rev. Lett. **48**, 1660 (1982).

29. Y. Aharonov and A. Casher, Phys. Rev. Lett. **53**, 319 (1984).

30. S.A. Werner, H. Kaiser, M. Arif, H.-C. Hu, and R. Berliner, Physica B&C **136**, 137 (1993).

31. S. Chu, in these proceedings.

32. Matter Wave Interferometry, a special review issue of Physica B&C **151**, (1988).

33. G. Timp, R.E. Behringer, D.M. Tennant and J.E. Cunningham, Phys. Rev. Lett. **69**, 1636 (1992).

FASTER THAN FOURIER

Michael Berry
H.H. Wills Physics Laboratory, Tyndall Avenue, Bristol BS8 1TL, U.K.

Written to celebrate the 60th Birthday of Yakir Aharonov: deep, quick, subtle.

ABSTRACT

Band-limited functions $f(x)$ can oscillate for arbitrarily long intervals arbitrarily faster than the highest frequency they contain. A class of integral representations exhibiting these 'superoscillations' is described, and by asymptotic analysis the origin of the phenomenon is shown to be complex saddles in frequency space. Computations confirm the existence of superoscillations. The price paid for superoscillations is that in the infinitely longer range where $f(x)$ oscillates conventionally its value is exponentially larger. For example, to reproduce Beethoveen's ninth symphony as superoscillations with a 1Hz bandwidth requires a signal $\exp\{10^{19}\}$ times stronger than with conventional oscillations.

1. Model for superoscillations

My purpose is to decribe some mathematics inspired by Yakir Aharonov during a visit to Bristol several years ago. He told me that it is possible for functions to oscillate faster than any of their Fourier components. This seemed unbelievable, even paradoxical; I had heard nothing like it before, and learned only recently of just one related paper[1] in the literature on Fourier analysis (see §4). Nevertheless, Aharonov and his colleagues had constructed such 'superoscillations' using quantum-mechanical arguments[2]. Here I will exhibit a large class of them, and use asymptotics and numerics to study their strange properties in detail.

Consider functions $f(x)$ whose spectrum of frequencies k is band-limited, say by $|k|\leq 1$, so that on a conventional view f should oscillate no faster than $\cos(x)$. But we wish f to be superoscillatory, that is to vary as $\cos(Kx)$, where K can be arbitrarily large, for an arbitrarily long interval in x. A representation that achieves this is

$$f(x,A,\delta)=\frac{1}{\delta\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}u\exp\{\mathrm{i}xk(u)\}\exp\left\{-\frac{1}{2\delta^2}(u-\mathrm{i}A)^2\right\} \qquad (1)$$

where the wavenumber function $k(u)$ is even, with $k(0)=1$ and $|k|\leq 1$ for real u, A is real and positive, and δ is small. Examples are

$$k_1(u)=\frac{1}{1+\frac{1}{2}u^2},\quad k_2(u)=\operatorname{sech}u,\quad k_3(u)=\exp\left\{-\tfrac{1}{2}u^2\right\},\quad k_4(u)=\cos u \qquad (2)$$

Aharonov's reasoning (he suggested Eq.(1) with k_4) was that when δ is small the second exponential would act like a 'complex delta-function' and so project out the value of the first exponential at u=iA. Thus f should vary as

$$f \approx \exp\{\mathrm{i}Kx\} \quad \text{where } K = k(\mathrm{i}A) \tag{3}$$

Under the conditions above Eq.(2), k increases from u=0 along the imaginary axis, so that K>1, (and for the given examples can be arbitrarily large), and so corresponds to superoscillations. What follows is a study of the small-δ asymptotics of the integral representing f. As well as justifying Aharonov's argument, this will dissolve the paradox posed by superoscillations, by showing that when x>O($1/\delta^2$) they get replaced by the expected cos(x), and f gets exponentially large.

2. Asymptotics

The aim is to get an asymptotic approximation for small δ to the integral defining f, Eq.(1), which is valid uniformly in x. To achieve this, it is convenient to define

$$\xi \equiv x\delta^2 \tag{4}$$

so that Eq.(1) can be written

$$f\left(\xi / \delta^2, A, \delta\right) = \frac{1}{\delta\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}u \, \exp\left\{-\frac{1}{\delta^2}\Phi(u,\xi,A)\right\} \quad \text{where } \Phi \equiv \tfrac{1}{2}(u - \mathrm{i}A)^2 - \mathrm{i}\xi\, k(u) \tag{5}$$

For small δ, f can now be approximated by the saddle-point method, that is by deforming the path of integration through saddles u_s of the exponent and replacing Φ by its quadratic approximation near u_s. f is dominated by the saddle with smallest ReΦ. Saddles, whose location depends on ξ (and also A) are defined by

$$\frac{\mathrm{d}\Phi}{\mathrm{d}u} = 0, \quad \text{i.e. } u_s = \mathrm{i}\left[\xi k'(u_s) + A\right] \tag{6}$$

Application of the saddle-point method now gives the main result:

$$f \approx \frac{\exp\left\{\mathrm{i}x\, k(u_s) - \frac{1}{2\delta^2}(u_s - \mathrm{i}A)^2\right\}}{\sqrt{1 - \mathrm{i}x\delta^2 k''(u_s)}} \tag{7}$$

To interpret this formula, it is necessary to understand the behaviour of the dominant saddle as ξ varies.

When ξ<<1, that is $x<<\delta^{-2}$, Eq.(6) gives $u_s \approx iA$, and (7) reduces to Eq.(3); this is the regime of superoscillations. When ξ>>1, that is $x>>\delta^{-2}$, the saddles are the zeros of $k'(u)$; assuming for simplicity that k has a single maximum at u=0 (as in the first three functions in Eq.(2)), this is the only real saddle, and (7) reduces to

$$f \approx \frac{1}{\delta\sqrt{x|k''(0)|}}\exp\left\{ix-\tfrac{1}{4}\pi\right\}\exp\left\{\frac{A^2}{2\delta^2}\right\} \tag{8}$$

This is the behaviour to be expected conventionally, that is on the basis of the frequency content of f; in the infinite range of validity of Eq.(8), f is O(exp{$A^2/2\delta^2$} and so exponentially amplified relative to the superoscillation regime.

As x increases, the saddle moves from iA to 0 along a curved track, illustrated in figure 1. This is the dominant saddle u_s; its track resembles figure 1 for all $k(u)$ of this type that I have studied. There are other solutions of Eq.(6), whose arrangement and motion are complicated and depend on the details of $k(u)$, but they are not dominant and so do not compromise the validity of Eq.(7) as the leading-order approximation to the integral defining f, Eq.(1).

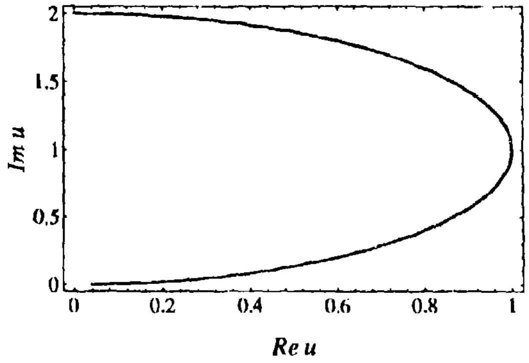


Figure 1. Track of leading saddle u_s as ξ increases from 0 to ∞, for the wavenumber function $k_5(u)$ in Eq.(10), for A=2 (the track is similar for any $k(u)$ with a single maximum)

In understanding the oscillations, it is helpful to study the local wavenumber, defined as

$$q(\xi) \equiv -\mathrm{Im}\frac{\partial\Phi\{u_s(\xi),\xi,A\}}{\partial\xi} = \mathrm{Re}\,k(u_s(\xi)) \tag{9}$$

As illustrated in figure 2, $q(\xi)$ decreases smoothly from k(iA) (which is real) to 1 as ξ increases. Note that the decrease is rapid (this is true for all $k(u)$ that I have studied). This has the important implication that to observe superoscillations it is necessary to keep ξ

much smaller than unity, and if we want to allow x to be large, in order to observe *many* superoscillations, δ must be correspondingly smaller, Eq.(4), and the exponential amplification in the regime of conventional oscillation, Eq.(8), will be correspondingly larger.

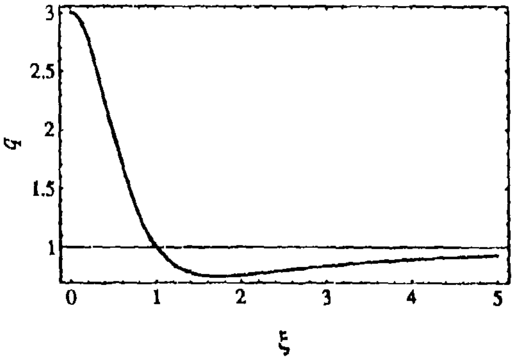


Figure 2. Local wavenumber $q(\xi)$, Eq.(9), for the $k_5(u)$ in Eq.(10), for A=2

None of the wavenumber functions in Eq.(2) gives an f whose integral representation can be evaluated exactly in terms of special functions. However, if we choose the wavenumber function

$$k_5(u) = 1 - \tfrac{1}{2}u^2 \qquad (10)$$

we can ensure that it is band-limited ($|k|<1$) by restricting the range of integration in Eq.(1) to $|u|\leq 2$. The resulting truncated integral is

$$f(x,A,\delta) = \frac{1}{\delta\sqrt{2\pi}}\int_{-2}^{2} \mathrm{d}u \exp\left\{\mathrm{i}x\left(1-\tfrac{1}{2}u^2\right)\right\}\exp\left\{-\frac{1}{2\delta^2}(u-\mathrm{i}A)^2\right\} \qquad (11)$$

which be expressed in terms of error functions:

$$\begin{aligned} f(x,A,\delta) &= \frac{1}{2\sqrt{1+\mathrm{i}x\delta^2}}\exp\left\{\frac{\mathrm{i}x\left(2+A^2+2\mathrm{i}x\delta^2\right)}{2\left(1+\mathrm{i}x\delta^2\right)}\right\}\times \\ &\times\left[\mathrm{erf}\left\{\frac{2+\mathrm{i}A+2\mathrm{i}x\delta^2}{\delta\sqrt{2+2\mathrm{i}x\delta^2}}\right\}+\mathrm{erf}\left\{\frac{2-\mathrm{i}A+2\mathrm{i}x\delta^2}{\delta\sqrt{2+2\mathrm{i}x\delta^2}}\right\}\right] \end{aligned} \qquad (12)$$

It is instructive to examine this in detail. The superoscillation wavenumber, Eq.(3), is

$$K = k_5(\mathrm{i}A) = 1 + \tfrac{1}{2}A^2 \tag{13}$$

There is a single saddle, at (figure 1)

$$u_s(\xi) = \frac{\mathrm{i}A}{1 + \mathrm{i}\xi} \tag{14}$$

and the local wavenumber is (figure 2)

$$q(\xi) = 1 + \frac{A^2(1-\xi^2)}{2(1+\xi^2)^2} \tag{15}$$

For this case, the saddle-point approximation, Eq.(7) gives

$$f(x,A,\delta) \sim \frac{1}{\sqrt{1+\mathrm{i}x\delta^2}}\exp\left\{\mathrm{i}x\left[1+\frac{A^2}{2(1+x^2\delta^4)}\right]\right\}\exp\left\{\frac{A^2\delta^2x^2}{2(1+x^2\delta^4)}\right\} \tag{16}$$

However, the asymptotics of (11) includes contributions from the end-points $u=\pm 2$ as well as the saddle u_s. This can be seen by realising that the steepest path between -2 and +2 runs from -2 to infinity in the negative half-plane, through u_s to infinity in the positive half-plane, and back to +2. The end-point contributions oscillate conventionally, with the wavenumber -1, so we must be sure that they do not mask the superoscillations that exist for small ξ. The condition for this is that the absolute value of the Gaussian in (11) must not exceed unity at the end-points. Thus

$$\exp\left\{\frac{A^2-4}{2\delta^2}\right\} \le 1, \qquad \text{i.e.} \qquad A \le 2 \tag{17}$$

(we include the equality because the end-point contribution is smaller than that from the saddle by a factor δ). Eq.(13) now implies that the maximum rate of superoscillation obtainable with this model is $K=3$. (It is worth remarking that $x=0$, $A=2$ lies on the anti-Stokes line for the error functions in Eq.(12), that is, where the exponential contribution from the saddle exchanges dominance with those from the end-points.)

The representation Eq.(1) does not have the form of a Fourier transform, namely (for a band-limited function)

$$f(x,A,\delta) = \int_{-1}^{1}\mathrm{d}q\,\exp\{\mathrm{i}xq\}\bar{f}(q) \tag{18}$$

It is however easy to cast it into this form. The transform $\bar{f}(q)$ depends on the inverse function of $k(u)$; this is multivalued, and the path of integration can be deformed into a loop around a cut extending along the real axis negatively from the branch point at $q=1$ (the ends of the loop are pinned to the cut, at $q=-1$ for k_5 and at the essential singularity $q=0$ for k_1, k_2, and k_3). Again there is a dominant saddle, which for small ξ lies at $q=K$, and the loop can be expanded to pass through this. All previous results can be reproduced in this way.

3. Numerics

The aim here is twofold: to compare the saddle-point approximation Eq.(7) with the exact integral (1), and to exhibit the superoscillations. I carried out computations of f for the wavenumber functions k_1, k_2, and k_3 (Eq.(2)), but will display results only for Re f (Im f is similar) for k_5 (Eq.(10)), with the truncated integral of Eq.(11), for which the results are very similar. The computations will be exhibited for the fastest superoscillations, namely $K=3$, that is $A=2$ (Eq.(17)), choosing $\delta=0.2$.

Figure 3 shows the results. The superoscillations for small x, with period $2\pi/3$, are shown on figure 3a, and figure 3b shows a range of x where there are conventional oscillations, with period more than 3 times greater (actually about 8.4 - cf. figure 2, where $\xi \sim 1.6$ corresponds to $x \sim 40$). In both cases, the approximation (in this case Eq.(16)) agrees well with the exact expression, Eq.(12). For example, the fractional error is 0.18 for $x=2$, and 2.8×10^{-18} for $x=42$. Note the enormous ratio of the sizes of f for large and small x; from Eq.(16), this can be estimated as $\exp(36)\sim10^{16}$ (the asymptotic ratio of Eq.(8) is not attained in figure 3b). The transition between the superoscillation and conventional regimes is clearly shown in figure 3c.

In these computations, the value $A=2$ is the largest for which the saddle dominates the end-points. The competition between contributions shows up most clearly at $x=0$, for which (12) gives

$$f(0,A,\delta) = \mathrm{Re\,erf}\left\{\frac{1}{\delta}\left(\sqrt{2}+\mathrm{i}\frac{A}{\sqrt{2}}\right)\right\} \tag{19}$$

For $A<2$, f is well approximated by the saddle contribution of unity, for $A>2$, the end-points dominate and f increases exponentially, Eq.(17), masking the superoscillations for small x. This is illustrated in figure 4. Even at the critical value $A=2$, that is, on the anti-Stokes line for the function (19), the exact value $f=0.945$ is close to the saddle-point value $f\sim1$.

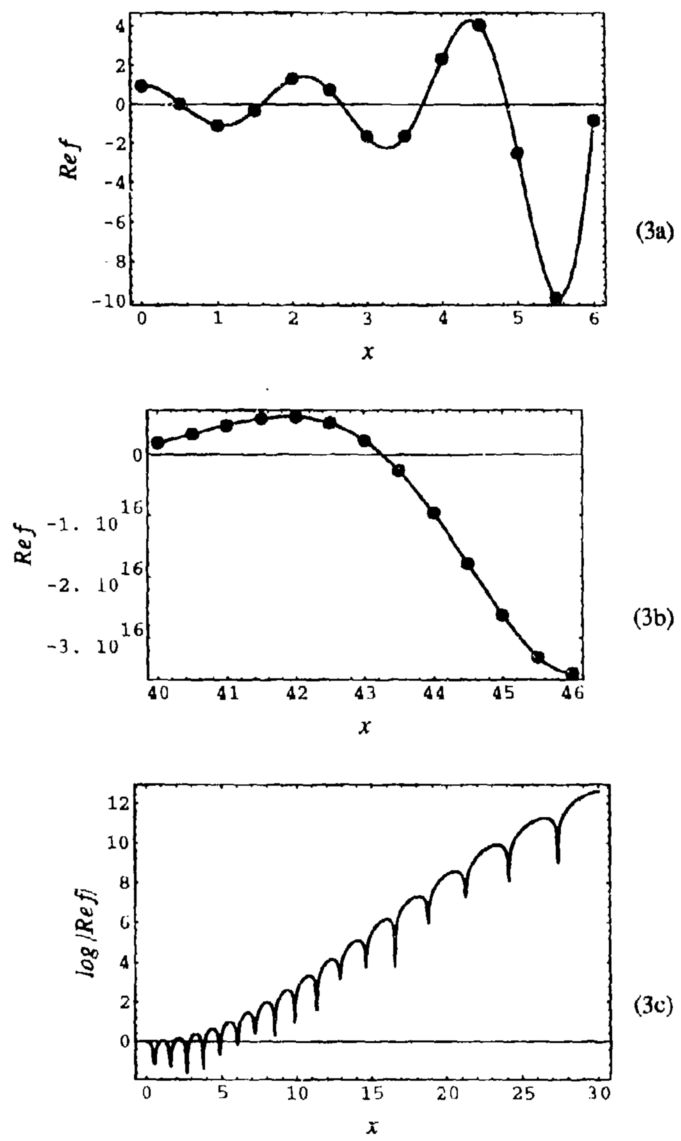


Figure 3. Computations of $f(x,2,0.2)$ for the truncated integral, Eq.(11), showing (a), superoscillations, and (b) conventional oscillations. Circles: exact expression, Eq.(12); full lines: saddle-point approximation, Eq.(16). In (c) the logarithms are base 10

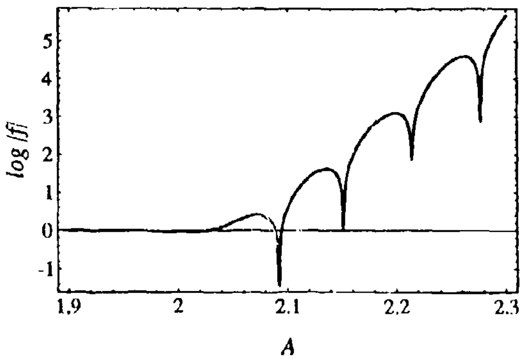


Figure 4. Computations of log $|f(0, A, 0.2)|$, Eq.(19), for the truncated integral Eq.(11); logarithms are base 10. Note the exponential growth after crossing the anti-Stokes line at A=2

4. Beethoven at 1Hz

Professor I. Daubechies has informed me that superoscillations are known in signal processing, in the context of oversampling. This is sampling a function faster than the Nyquist rate, i.e. at points $x=n\pi$ where the function is band-limited by $|k|\leq 1$. If a function is oversampled in a finite range, extrapolation outside this range is exponentially unstable[2]. She quotes B. Logan as saying that it is possible in principle to design a bandlimited signal with a bandwidth of 1Hz that would reproduce Beethoven's ninth symphony exactly. With the superoscillatory functions described in this paper it is possible to give an explicit recipe for constructing this signal, as I now explain.

We require superoscillations for the duration T (~4000s) of the symphony. Therefore the desired signal $B(t)$ can be represented as periodic outside this interval, namely

$$B(t)=\sum_{-N}^{N} B_n \exp\left\{\mathrm{i}\frac{2\pi nt}{T}\right\} \qquad (20)$$

Here N is the order of the Fourier component corresponding to the highest frequency $\nu_{\max}\equiv N/T$ (~20kHz) it is desired to reproduce.

To approximate this with a signal band-limited by frequency ν_0 (1Hz) we make the replacement

$$\exp\left\{\mathrm{i}\frac{2\pi nt}{T}\right\}\rightarrow \Phi_n(t) \qquad (21)$$

where (cf.Eq.(1)) Φ_n is the superoscillatory function

$$\Phi_n(t) \equiv \frac{1}{\delta_n\sqrt{2\pi}} \int_{-\infty}^{\infty} du \exp\{i2\pi t v(u)\} \exp\left\{-\frac{1}{2\delta_n^2}(u - iA_n)^2\right\} \tag{22}$$

Here the frequency function $v(u)$ never exceeds (for real u) its band-limited value $v(0) \equiv v_0$, and A_n and δ_n will now be determined by the requirement that Φ_n superoscillates with frequency n/T for time T.

The superoscillation frequency of $\Phi_n(t)$ is $v(iA_n)$ (cf. Eq.(3)). Thus from Eq.(21) A_n must satisfy

$$v(iA_n) = \frac{n}{T} \tag{23}$$

We fix δ_n by requiring that the superoscillations are maintained for time T, in the sense that the replacement of Eq.(21) remains a good approximation. For this we require the next correction to the superoscillatory exponential that $\Phi_n(t)$ represents. Expanding the saddle-point approximation to Eq.(22) (analogous to Eq.(7)) for small t, we find

$$\Phi_n(t) \approx \exp\left\{i\frac{2\pi nt}{T}\right\} \exp\left\{2\pi^2\delta_n^2\left[-v'^2(iA_n)\right]t^2\right\} \tag{24}$$

The second factor is an increasing exponential, because $v'(iA_n)$ is imaginary, and must remain close to unity for $0<t<T$. Thus

$$\delta_n << \left[2\pi|v'(iA_n)|T\right]^{-1} \tag{25}$$

Choosing A_n and δ_n as in Eqs.(23) and (25) guarantees that the signal $B_n(t)$, with its frequencies up to $v_{\max}$, will be imitated for time T. When $t>T$ the imitation will grow rapidly in strength, and eventually, that is when it is oscillating at the frequency v_0 corresponding to its Fourier content, it will acquire an amplification factor corresponding to its largest Fourier component $n=N$. An argument analogous to that leading to Eq.(8) gives this factor as

$$F = \exp\left\{\frac{A_N^2}{2\delta_N^2}\right\} >> \exp\left\{A_N^2\pi^2T^2|v_N'(iA_N)|^2\right\} \tag{26}$$

with A_N determined by Eq.(23) with the right-hand side set equal to $v_{\max}$.

Let us calculate this amplification for the model frequency function

$$v(u) = v_0\exp\{-u^2\} \tag{27}$$

(cf. $k_3(u)$ in Eq.(2)). We find

$$A_N^2 = \log\left\{\frac{\nu_{\max}}{\nu_0}\right\} \tag{28}$$

and hence, from Eq.(26),

$$F >> \exp\left\{4\pi^2 \log^2\left(\frac{\nu_{\max}}{\nu_0}\right)\nu_{\max}^2 T^2\right\} \tag{29}$$

For Beethoven's ninth symphony this gives

$$F >> \exp\left\{10^{19}\right\} \tag{30}$$

This amplification will not be achieved until a time t_F, which can be estimated by the argument preceding Eq.(8) as

$$t_F \sim \left[\nu_0 \delta_N^2\right]^{-1} \sim \frac{\nu_{\max}^2 T^2}{\nu_0} \sim 10^8 \text{years} \tag{31}$$

Other choices for $\nu(u)$ give similar expressions and numerical estimates.

The estimate of Eq.(30) indicates that to reproduce music as superoscillations requires a signal with so much energy as to be hopelessly impractable, but more modest bandwidth compression might be feasible.

5. Concluding remarks

Aharonov's discovery, elaborated here, could have applications in several branches of physics. One possibility is the use of superoscillations for bandwidth compression as discussed in §4. Another example, also in signal processing, concerns the observation of oscillations faster than those expected on the basis of applied or inferred filters. These would conventionally be interpreted as high frequencies leaking through imperfect filters, but the arguments presented here show that the phenomenon could have a quite different origin, namely superoscillations compatible with perfect filtering.

Perhaps more interesting are the possible applications of superoscillatory functions of two variables, representing images. One envisages new forms of microscopy, in which structures much smaller than the wavelength λ would be resolved by representing them as superoscillations. (This is different from conventional superresolution, which is based on the fact that Fourier components larger than $2\pi/\lambda$ can be present in the field near the surface of an object, but decay exponentially away from the object because the wavenumber in the perpendicular direction is imaginary. With superoscillations, the larger Fourier components are not present.)

Superoscillauons can probably exist in random functions $f(x)$: arbitrarily long intervals, in which f is exponentially small relative to elsewhere, could superoscillate. Consider how this might be achieved. If f is Gauss-distributed, its statistics are completely described by its autocorrelation function, which by the Wiener-Khinchin theorem is the Fourier transform of the power spectrum $S(q)$ of f. Even if f is band-limited, it ought to be possible to choose $S(q)$ with analytic structure (saddles with Re q >1, etc.) such that the autocorrelation superoscillates as it falls from its initial value. This idea is worth pursuing.

On the purely mathematical side, it is clear that superoscillations carry a price: the function is exponentially smaller than in the regime of conventional oscillations, with the exponent increasing with the size of the interval of superoscillations. We have seen examples of this, but there ought to be a general theorem (perhaps based on a version of the uncertainty principle).

6. Acknowledgments

I thank Professors Jeeva Anandan and John Safko for arranging, and generously supporting my participation in, Yakir Aharonov's birthday meeting, and inviting me to write this paper, and Professor Ingrid Daubechies for suggestions leading to the calculation of §4.

7. Reference

1. H.J. Landau, IEEE Trans. Inf. Theory. **IT-32** (1986) 464
2. Y. Aharonov, J. Anandan, S. Popescu and L. Vaidman, *Phys.Rev.Lett.* **64** (1990) 2965.

BERRY'S PHASE, MESOSCOPIC CONDUCTIVITY AND LOCAL FORCES

ADY STERN
Lyman Laboratory of Physics, Harvard University
Cambridge, Ma. 02138

ABSTRACT

A ring in a magnetic field whose direction varies in space is considered. It is shown that the Berry phase accumulated by the spins of electrons encircling the ring affects the conductance of the ring in a way similar to the Aharonov–Bohm effect. A time–dependent Berry phase is shown to induce a classical motive force in the ring. The condition for adiabaticity is studied, as well as deviations from that condition. The relation to spin–orbit coupling is discussed.

1. Introduction

This paper studies an effect of the geometric (Berry's) phase[1][2]on electronic transport in mesoscopic and macroscopic systems.[3]The reader might be somewhat surprized by the order in which the subject is presented below. To some extent, that order resembles a presentation of the theory of electromagnetism, ***but in reversed order***. A study of electromagnetism usually starts with a description of Coulomb's and Lorenz' forces. Then, the concept of potentials is presented, as a tool for calculating forces and fields. And finally, the special role given by Quantum Mechanics to vector potentials, as geometric "phase shifters", is introduced, and the non–local nature of Quantum Mechanics is revealed. This paper, however, like many other studies of Berry's phase, starts with an investigation of a quantum mechanical geometric phase. In the case discussed below, the phase is accumulated by an electron's spin moving in a space–dependent magnetic field. Then, this effect is put in terms of a vector potential. And finally, the effect of this vector potential on the classical dynamics is revealed.

2. A conducting ring in a space–dependent magnetic field

The simplest example that illustrates the concept of Berry's phase is that of a spin-$\frac{1}{2}$ that follows adiabatically a magnetic field whose direction varies in time. When the magnetic field returns to its initial direction, the spin wave function is found to have acquired a geometric phase factor, given by half the solid angle subtended by the magnetic field during its variation. This phase can be regarded as induced by a geometric flux, similar to the phase shift induced by an electromagnetic flux in the Aharonov–Bohm effect.(1)(2)

Motivated by this similarity between the fluxes, we turn to investigate Berry's phase analogies to two physical effects involving an electromagnetic flux: the induction of current in a conducting ring by a time dependent electro-magnetic flux (through Faraday's law), and the effect of time–independent flux on the conductivity of a mesoscopic ring (through the Aharonov–Bohm effect).(4)(5)In these analogies, the electron's spin plays the role played by the electric charge in the electromagnetic effects. Another analogy, introducing persistent currents induced by Berry's phase in ballistic rings, was recently discovered in an instructive work of Loss, Goldbart and Balatsky(6). In the following paragraph we define a thought experiment in which electrons in a mesoscopic conducting ring follow adiabatically a magnetic field whose direction varies spatially, and thus accumulate Berry's phase. By mapping that phase onto an effective vector potential, we show that when the phase is time–independent, it affects the ring's conductance. When the phase varies in time, it induces a current in the ring. By discussing the analogies to the electromagentic phenomena, we point out that the effect of a time–independent geometric flux is observable only in mesoscopic rings, while the effect of a time–dependent geometric flux should be observed also in macroscopic rings, i.e., it does not depend on phase coherence. Since the adiabatic approximation is crucial for this discussion, we examine the conditions for its validity, and its dependence on the disorder in the ring. We also comment on the remnants of the geometric phase in the non–adiabatic limit, and on the relations of these effects to spin–orbit coupling. While for practical reasons our discussion is concentrated on the electric properties of the ring, we nevertheless stress that *the electric charge of the electron plays no role in our analysis.* Our results stem from the Zeeman interaction, and are therefore valid for all spin-$\frac{1}{2}$ particles, irrespective of their charge.

We consider a quasi-one dimensional ring, whose radius is a. The ring lies in the $x \otimes y$ plane, and its center is in the origin. A non–uniform magnetic field is applied on the ring in the following way: first, a magnetic field B_ϕ tangent to the ring is induced by a current carrying wire lying along the z–axis. Second, a uniform field, B_z, is applied on the system, parallel to the z-axiz. Adopting a cylindrical coordinate system, the total magnetic field has a component B_ϕ created by the wire at the $\hat{\phi}$ direction, and a component B_z at the $\hat{z}$ direction. Along the ring, the magnitude of the field is constant, but the direction varies. In fact, it follows a cone shaped path, where the angle between the cone and the z-axis, denoted by α, satisfies $\tan\alpha = \frac{B_\phi}{B_z}$ (See Fig. 1). The spin of an electron that slowly encircles the ring is then expected to follow the direction of the magnetic field and thus accumulate a geometrical phase of

$$\Omega_g^{\uparrow(\downarrow)} \equiv \pi(1 \pm \cos\alpha) \tag{1}$$

i.e., half the solid angle subtended by the the magnetic field it goes through (The ↑, + and ↓, – refer to the spin being parallel and anti–parallel to the field, respectively). The angle α is determined by the current through the wire and by the uniform field along the z direction, both of which we regard as the controlled variables in the experiment.

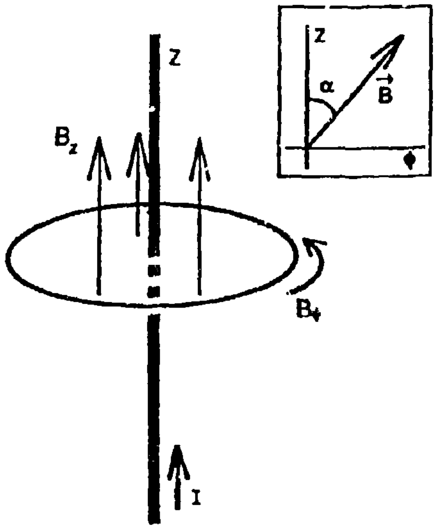


Figure 1: The physical problem considered. A ring is put in a uniform external magnetic field B_z, and a tangential magnetic field B_ϕ created by the current carrying wire. The ratio between the two fields define the angle α.

3. The adiabatic approximation

Our discussion of the above described thought experiment involves several parts. In this section we use the Born–Oppenheimer approach in order to separate the Hamiltonian of the system into two parts, one (the adiabatic part) in which the spin follows adiabatically the direction of the magnetic field, and one (the non-adiabatic part) which is purely non-diagonal with respect to the eigenstates of the adiabatic part. We show that the adiabatic part includes a geometric vector potential that couples to the electron's spin. Assuming that the ring is one dimensional, its Hamiltonian is

$$H = \frac{\Pi^2}{2M} + V(\phi) - \mu\vec{B}(\phi)\cdot\vec{\sigma} \qquad (2)$$

where $\Pi = -\frac{i}{a}\frac{d}{d\phi} - \frac{eB_z\pi a}{hc}$ is the generalized momentum (a system of units where $\hbar = 1$ is utilized), $V(\phi)$ is the impurity potential along the ring, μ is the magnetic moment, M is the mass of an electron, and $\vec{\sigma}$ is the Pauli matrices vector. Attempting to discuss the adiabatic limit, we diagonalize the spin dependent part of the Hamiltonian, treating the angle ϕ as a parameter. Denoting the two eigenstates by $|\uparrow(\phi)\rangle$ ($|\downarrow(\phi)\rangle$), corresponding to the spin being parallel (anti-parallel) to the magnetic field, we get

$$|\uparrow(\phi)\rangle = \begin{pmatrix} i\cos\frac{\alpha}{2}e^{-i\phi} \\ -\sin\frac{\alpha}{2} \end{pmatrix} \qquad \text{and} \qquad |\downarrow(\phi)\rangle = \begin{pmatrix} i\sin\frac{\alpha}{2}e^{-i\phi} \\ \cos\frac{\alpha}{2} \end{pmatrix} \tag{3}$$

The corresponding eigenvalues are $\mp\mu B$ where $B \equiv \sqrt{B_\phi^2 + B_z^2}$. Defining now $|\phi\rangle$ as the cigenstate of the operator $e^{i\hat{\phi}}$, the two sets of states $\left\{|\uparrow(\phi)\rangle \otimes |\phi\rangle \;|0 \leq \phi < 2\pi\right\}$ and $\left\{|\downarrow(\phi)\rangle \otimes |\phi\rangle \;|0 \leq \phi < 2\pi\right\}$ constitute together a basis of the Hilbert space of the Hamiltonian (2) . Each one of these sets span a subspace in which the spin is either parallel or anti-parallel to the magnetic field. The impurity potential is spin independent, and hence, it is diagonal in that basis. However, the kinetic part of the Hamiltonian has matrix elements that connect states within the subspaces defined above, as well as matrix elements that connect states of different subspaces, i.e., induce spin-flips. A simple calculation shows that the matrix elements connecting states within the first sub-space are,

$$\langle\uparrow(\phi)| \otimes \langle\phi|\frac{\Pi^2}{2M}|\phi'\rangle \otimes |\uparrow(\phi')\rangle = \langle\phi|\frac{[\Pi - \frac{1}{2\pi a}\Omega_g^\uparrow]^2}{2M}|\phi'\rangle \tag{4}$$

The corresponding matrix element in the second subspace has $\Omega_g^\downarrow$ rather than $\Omega_g^\uparrow$.

These matrix elements demonstrate that within the adiabatic approximation, the spatial variation of the magnetic field induces a vector potential(7)whose magnitude is independent of the electron's charge, but is rather determined by the direction of the spin being parallel or anti-parallel to the field. Following the method outlined recently by Aharonov *et.al.*,(8)we construct an operator A_g in such a way that the operator $\frac{1}{2M}[\Pi - A_g]^2$ has diagonal matrix elements given by Eq. (4), and does not have any elements connecting states with opposite spin direction. A simple calculation shows that

$$A_g = \frac{1}{2a}\sin\alpha[\cos\alpha\vec{\sigma}\cdot\hat{\phi} - \sin\alpha\sigma_z] \tag{5}$$

Note that $A_g^2 = \frac{1}{4a^2}\sin^2\alpha$ is a c-number, and A_g has non-zero matrix elements only between states of opposite spin directions. Consequently, the separation of the Hamiltonian to an adiabatic part, H_0, and a purely non-adiabatic part, H_1, is given by

$$H_0 = \frac{[\Pi - A_g]^2}{2M} + V(\phi) - \mu\vec{B}(\phi)\cdot\vec{\sigma} + \frac{1}{8Ma^2}\sin^2\alpha \tag{6}$$

and

$$H_1 = \frac{1}{2M}[(\Pi - A_g)A_g + A_g(\Pi - A_g)] \tag{7}$$

By construction, H_0 has a set of eigenstates $|n,\uparrow\rangle = |\uparrow(\phi)\rangle \otimes \psi_n^\uparrow(\phi)$ in which the spin is parallel to the field, and a set of eigenstates $|n,\downarrow\rangle = |\downarrow(\phi)\rangle \otimes \psi_n^\downarrow(\phi)$ in which the spin is anti-parallel to the field. The wave functions $\psi_n^\uparrow(\phi)$ and $\psi_n^\downarrow(\phi)$ satisfy the Schroedinger equations $H_0^{\uparrow(\downarrow)}\psi_n^{\uparrow(\downarrow)} = E_n^{\uparrow(\downarrow)}\psi_n^{\uparrow(\downarrow)}$, where the Hamiltonians $H_0^{\uparrow(\downarrow)}$ are given by,

$$H_0^{\uparrow(\downarrow)} \equiv \left\{\frac{1}{2M}\left[\Pi - \frac{1}{2\pi a}\Omega_g^{\uparrow(\downarrow)}\right]^2 + V(\phi) \mp \mu B + \frac{1}{8Ma^2}\sin^2\alpha\right\} \tag{8}$$

Each of these Hamiltonians is a projection of the full Hamiltonian onto one of the subspaces defined above. The meaning of the induced vector potential becomes clearer when one considers space translation transformations. The momentum operator, $\hat{p}_\phi$, is, of course, the generator of such a transformation, i.e., for any state Ψ, $\langle\phi|e^{-i\frac{\hat{p}_\phi\phi_0}{\hbar}}|\Psi\rangle = \langle\phi+\phi_0|\Psi\rangle$. In such a transformation, the electron is translated spatially, but the direction of the spin is kept constant. On the contrary, the generalized momentum appearing in the adiabatic Hamiltonian, $\Pi - A$ is the generator of a different translation transformation, a transformation in which the electron is translated spatially, *and the direction of the spin follows the direction of the field.*

We conclude this section by emphasizing its main conclusion: Under conditions in which the adiabatic approximation is valid, namely, H_1 can be disregarded, the ring can be viewed as composed of two uncoupled electron gases. Those gases are subject to the effect of different geometric vector potentials and opposite constant potential energy, originating from the Zeeman interaction. They are also subject to the effect of identical electromagnetic flux $B_z\pi a^2$ and identical impurity potential. Each of the two gases obviously does not have a spin degeneracy.

4. Non–local and local effects of the geometric flux on electronic transport

In the next part of the discussion we assume that the magnetic field is strong enough for the adiabatic limit to be applicable. The discussion of the precise meaning of "strong enough" is postponed to the next section. Assuming that the Zeeman energy μB is smaller than the Fermi energy, ϵ_F, our ring consists of the two uncoupled electron gases described above. The electric conductance of the ring is then the sum of the conductances of the two gases. As discussed extensively in recent years,(4)(8) the conductance of a mesoscopic ring depends on a magnetic flux threading the ring, through the Aharonov–Bohm effect. For rings in the diffusive regime, the flux dependence of the conductance is manifested in two different contexts, namely, the average conductance of an ensemble of macroscopically identical rings and the sample–specific fluctuations. The flux–dependent part of the average conductance was calculated by Al'tshuler, Aronov and Spivak(9), and shown to be,

$$\delta\sigma = -\frac{e^2a}{\Gamma}\frac{\sinh(\Gamma)}{\cosh(\Gamma)-\cos(\frac{4\pi\phi}{\phi_0})} \tag{9}$$

where ϕ is the flux threading the ring, $\Gamma \equiv \frac{2\pi a}{l_\phi}$ and l_ϕ is the phase breaking length. Adjusted for our purposes, this expression is written for one spin direction. In the configuration we discuss, the flux threading the sample is a sum of an electromagnetic flux $\phi_{em} = B_z\pi a^2$, and the geometric flux $\phi_g = \frac{\phi_0}{2}(1\pm\cos\alpha)$, where the ± refers to electrons whose spin is parallel (anti–parallel) to the field. It should be noted here that the sum of the two geometric fluxes corresponding to the two gases equals a flux quantum. This stems from the fact that the sum of the geometric phases accumulated by the two spin directions is 2π. Since all properties of the ring are periodic with respect to one flux quantum, one can view the two electron gases as subject to the influence of geometric fluxes of equal magnitude and opposite directions. The total quantum correction to the conductivity is given by

$$\delta\sigma = -\frac{e^2 a}{\Gamma}\Big[\frac{\sinh(\Gamma)}{\cosh(\Gamma) - \cos(\frac{4\pi(\phi_{em}+\phi_g)}{\phi_0})} + \frac{\sinh(\Gamma)}{\cosh(\Gamma) - \cos(\frac{4\pi(\phi_{em}-\phi_g)}{\phi_0})}\Big] \tag{10}$$

This quantum correction to the classical Drude conductance results from interference of pairs of time–reversed paths.(10)The flux dependence stems from the phases accumulated by those paths that encircle the ring. When $\Gamma > 1$, the interference of long paths that encircle the ring more than once is exponentially suppressed, and the flux dependent correction to the conductivity can be approximated by

$$\delta\sigma = -\frac{2e^2 a}{\Gamma} e^{-\Gamma} \cos(\frac{4\pi\phi_{em}}{\phi_0}) \cos(\frac{4\pi\phi_g}{\phi_0}) \tag{11}$$

Then, the $\frac{\phi_0}{2}$ periodicity of the Aharonov–Bohm oscillations of the conductance is multiplied here by a geometrical factor, $\cos(\frac{4\pi\phi_g}{\phi_0})$. Note that the difference between the Fermi wavelengths of the two spin directions is not reflected in the expressions above, since the quantum correction to the conductivity is independent of $k_F l_{el}$.

The effect of the geometric flux on the sample-specific fluctuations of the conductance is best understood when the periodicity of those oscillations with respect to B_z is considered. In the absence of geometric flux ($B_\phi = 0$), the ϕ_0 flux periodicity yields a field periodicity of $\Delta B_z = \frac{\phi_0}{\pi a^2}$, irrespective of the spin direction. In the presence of geometric flux, a variation of B_z varies both the electromagnetic and the geometric fluxes. Thus the periodicity with respect to B_z is changed, and is no more independent of the spin direction. Specifically, when $B_z \ll B_\phi$ (i.e., $\alpha \to \frac{\pi}{2}$), the geometrical flux is approximately $\frac{\phi_0 B_z}{2B_\phi}$, and the B_z period becomes,

$$\Delta B_z = \frac{\phi_0}{\pi a^2 \pm \frac{\phi_0}{2B_\phi}} \tag{12}$$

where the +(–) sign refers to the spin being parallel (anti–parallel) to the field. The magnitude of the sample–specific fluctuations is not affected by the geometric flux, i.e., it is of the order of $\frac{e^2}{\hbar}$.

Eqs. (10) – (12) summarize our predictions for the effect of Berry's phase on the conductivity of a mesoscopic ring. We now turn to discuss the case of a time–dependent geometric flux, and, in particular, the currents it induces in the ring. We consider the case in which the tangential magnetic field is $B_\phi = B_\phi^0 \cos\omega t$. In order to avoid, at this stage, the complications involved in the analysis of the adiabatic condition for that case, we limit ourselves to the case in which the electron gas is completerly spin–polarized. This is realized when $\epsilon_F + \omega \ll \mu B_z$, i.e., the electron gas is spin–polarized, and an absorbtion of an energy quantum $\hbar\omega$ still does not allow electrons to flip their spins. For semi–conducting rings, this condition may be fulfilled at fields of the order of 1 Tesla. By passing, we note that another way to realize a completely spin polarized electron gas is by an injection of spin polarized electrons through a ferromagnetic–metallic interface.(11)Under the assumption of complete spin polarization, the electron gas in the ring is subject to the effect of a time-dependent geometrical flux

$$\phi_g = \frac{\phi_0}{2}(1 + \cos\alpha(t)) = \frac{\phi_0}{2}\left\{1 + \frac{1}{\sqrt{1 + (\frac{B_\phi^0}{B_z}\cos\omega t)^2}}\right\}.$$

Consequently, this gas is subject to a motive force ϵ, given by $\epsilon = -\frac{d\phi_g}{dt}$, and this motive force induces a current in the ring, according to Ohm's law. Assuming that $B_\phi^0 \ll B_z$, the motive force induced by the time dependence of ϕ_g is

$$\epsilon = -\frac{\omega\phi_0(B_\phi^0/B_z)^2}{2}\sin 2\omega t \tag{13}$$

The frequency of the induced current is twice as large as that of B_ϕ, so that it can experimentally be distinguished from currents induced due to the wire being not exactly perpendicular to the ring. For $B_z = 1$ Tesla, $B_\phi = 0.2$ Tesla and $\omega = 1$ GHz, this motive force has an amplitude of 10^{-7} Volts. Similarly to the electromotive force, the geometric motive force can be amplified if the ring is replaced by a solenoid.

There are a few points that should be stressed regarding the case of a time dependent geometrical flux. Firstly, contrary to the effect of a time-independent flux, the time dependent geometric flux exerts a force on the electron, [12]similar to the electric force exerted by a time–dependent electromagnetic flux. Thus, similar to the obseravtion of currents induced due to Faraday's law, the observation of currents induced by the geometric flux *does not depend on the electron phase being coherent along the ring.* Those currents should be observed in macroscopic rings, as well as in mesoscopic ones. In fact, the force accelerating the electrons in the case of a time dependent geometric flux is classical.[13]Secondly, the motive force induced in the ring is not electric, since if the electrons were replaced by neutrons, the picture would not have changed. The field, given by the derivative of the vector potential with respect to the time, does not couple to the electric charge, but rather to the direction of the spin. Thirdly, the origin of the motive force exerted on the electron can be understood by noting that in our symmetrical structure the sum of the orbital and spinor angular momenta in the z direction is time-independent even when the angle α is time dependent. Thus, a change in α transfers angular momentum from the spin to the orbital motion of the electron. A more general analysis of this force, from the point of view of classical equations of motion is given in Ref. (13).

So far we have discussed the currents induced by the geometric motive force only in the case of complete spin-polarization of the electrons. However, the flux, motive force and current all depend on the direction of the spin. Therefore, if the ring includes two electron gases with opposite spin directions, the currents induced in the two gases are opposite in direction, and the net current is proportional to the difference between the conductances of the two electron gases in the ring. Such a difference arises from the $2\mu B$ difference between the kinetic energy of electrons in the Fermi levels of the two electron gases.

5. Conditions for the validity of the adiabatic approximation

In this section we analyze the conditions under which the non–adiabatic part of the Hamiltonian, H_1, can be disregarded. Our discussion concentrates on the time–independent magnetic field and on the non–local effects it induces in the electronic transport of the ring. We start the discussion by considering the ballistic case, where $V(\phi) = 0$, a case for which the full Hamiltonian can be exactly diagonalized. For a ballistic ring the eigenstates of both $H_0^{\uparrow(\downarrow)}$ are given by $\psi_n^{\uparrow(\downarrow)}(\phi) = \frac{1}{\sqrt{2\pi}}e^{in\phi}$. The matrix elements of H_1 connect only states of opposite spin direction and *identical* spatial wave function. They are given by $\langle m, \downarrow |H_1| n, \uparrow\rangle = -\frac{(2n'-1)}{4Ma^2}\sin\alpha\,\delta_{n,m}$, where $n' \equiv n - \frac{eB_z a^2}{2c}$. Consequently, the *exact* eigenstates of the full Hamiltonian (2) are given by

$$|n, \uparrow(\phi)\rangle = e^{in\phi}\begin{pmatrix} i\cos\frac{\gamma}{2}e^{-i\phi} \\ -\sin\frac{\gamma}{2} \end{pmatrix} \qquad \text{and} \qquad |n, \downarrow(\phi)\rangle = e^{in\phi}\begin{pmatrix} i\sin\frac{\gamma}{2}e^{-i\phi} \\ \cos\frac{\gamma}{2} \end{pmatrix} \tag{14}$$

where γ is implicitly given by

$$\cot\gamma = \cot\alpha + \frac{\hbar^2(2n'-1)}{4Ma^2\mu B\sin\alpha} \tag{15}$$

The corresponding eigenvalues for $|n, \uparrow\rangle$ and $|n, \downarrow\rangle$ are

$$E(n) = \frac{\hbar^2 n^2}{2Ma^2} - \frac{\hbar^2(2n'-1)}{4Ma^2}(1 \pm \cos\gamma) \mp \mu B\cos(\gamma - \alpha) \tag{16}$$

The adiabatic approximation taken in the previous sections amounts to approximating $\gamma = \alpha$ for eigenstates for which the spin direction is parallel to the magnetic field and $\gamma = \alpha + \pi$ for eigenstates for which the spin direction is anti-parallel to the field. As seen from Eq. (15) , the adiabatic approximation is valid, for a ballistic ring, when $\mu B \gg \frac{n'}{Ma^2}$. The physical meaning of this result is better understood when noting that $\frac{Ma^2}{n'}$ is the time it takes an electron whose momentum is $\frac{n'}{a}$ to encircle the ring. The adiabatic approximation is then valid when this time is much longer than the precession time of the spin. Since our main interest is in the validity of the adiabatic approximation for electrons at the Fermi level, where $\frac{n'}{Ma} = v_F$ is the Fermi velocity, the condition for the adiabatic approximation to hold is,

$$\frac{\mu B a}{v_F} \gg 1 \tag{17}$$

The exact solubility of the ballistic case allows for a detailed analysis of deviations from the adiabatic limit. This analysis is given in the next section.

In the presence of impurity potential, the eigenstates of $H_0^{\uparrow(\downarrow)}$ are not eigenstates of the momentum operator Π, and therefore H_1 couples each eigenstate $|n, \uparrow\rangle$ to a continuum of states $|m, \downarrow\rangle$ (and vice versa). Due to that coupling, each adiabatic eigenstate acquires a finite lifetime, τ. We now calculate this lifetime perturbatively using the diagrammatic impurity technique.[8] According to Fermi's golden rule, the scattering time from a state $|n,\uparrow\rangle$ due to the perturbation H_1 is

$$\frac{1}{\tau}(|n,\uparrow\rangle) = \frac{2\pi}{\hbar}\sum_{|m,\downarrow\rangle} |\langle n,\uparrow|H_1|m,\downarrow\rangle|^2\delta(E_n^\uparrow - E_m^\downarrow) \tag{18}$$

(Note that H_1 is purely non-diagonal in spin states). While this lifetime is a meaningful quantity for a given ring with a given impurity configuration, it is not suitable for impurity averaging – one cannot identify a state $|n,\uparrow\rangle$ in two rings of different impurity configurations. Therefore, we define the average lifetime for a state with energy E, $\frac{1}{\tau}(E)$, as the average of $\frac{1}{\tau}(|n,\uparrow\rangle)$ over all states $|n,\uparrow\rangle$ with energy E:

$$\frac{1}{\tau}(E) = \frac{2\pi}{\hbar}\sum_{|n,\uparrow\rangle}\frac{1}{\nu^\uparrow(E)}\delta(E - E_n^\uparrow)\sum_{|m,\downarrow\rangle} |\langle n,\uparrow|H_1|m,\downarrow\rangle|^2\delta(E - E_m^\downarrow) \tag{19}$$

where $\nu^\uparrow$ is the density of states with the spin parallel parallel to the field. The corresponding expression for the lifetime of a state $|m,\downarrow\rangle$ has $\nu^\downarrow$ rather than $\nu^\uparrow$. Next, we examine the perturbation H_1. This perturbation is a product of two operators. The first, A_s, flips the spin state from being parallel to the field to being antiparallel, but does not affect the spatial wavefunction $\psi_n^\uparrow$. The second, $\frac{1}{M}(\Pi - A_s)$, is the projection of the velocity operator onto the spin–diagonal subspace. Thus, the average lifetime for a state with energy E is

$$\begin{aligned}\frac{1}{\tau}(E) = &\ \frac{2\pi}{\hbar\nu^\uparrow(E)}\frac{\sin^2\alpha}{(2a)^2}\textstyle\sum_{|n,\uparrow\rangle}\sum_{|m,\downarrow\rangle}\delta(E - E_n^\uparrow)\delta(E - E_m^\downarrow)\left|\int d\phi\psi_n^\uparrow(\phi)\hat{v}\psi_m^\downarrow(\phi)\right|^2 \\ = &\ \frac{2\pi}{\hbar\nu^\uparrow(E)}\frac{\sin^2\alpha}{(2a)^2}\textstyle\sum_{|\psi_n\rangle}\langle\psi_n|\hat{v}\delta(E - H_0^\uparrow)\hat{v}\delta(E - H_0^\downarrow)|\psi_n\rangle\end{aligned} \tag{20}$$

where $\hat{v} \equiv \frac{1}{M}(-i\frac{d}{d\phi} - \frac{eB_z\pi a}{2c})$. Note that the second line of Eq. (20) is all expressed in terms of single particle *spinless* operators and wave functions. There are two differences between the two spinless Hamiltonians $H_0^\uparrow, H_0^\downarrow$. First, they differ in the sign of the Zeeman energy. Second, they differ in the value of the geometric flux. If the second difference is disregarded for the moment, then the Zeeman energy difference can be absorbed in the energy arguments of the δ-functions. When this is done the two Hamiltonians become identical, but the energy arguments in the two δ-functions differ in $2\mu B$. Then, Eq. (20) strongly resembles Kubo's formula for the *ac* conductivity,

$$\sigma_{ac}(\omega) = \frac{4\pi e^2}{\hbar}\sum_{\psi_n}\langle\psi_n|\delta(\epsilon_F + \hbar\omega - H_0)\hat{v}\delta(\epsilon_F - H_0)\hat{v}|\psi_n\rangle \tag{21}$$

Thus, one might expect that under conditions in which the flux sensitivity of $H_0^\uparrow, H_0^\downarrow$ can be neglected, the average life-time $\frac{1}{\tau}(E)$ is proportional to the *ac* Kubo conductivity, at frequency $2\mu B$. This neglect can be expected to be valid up to a leading order in $\frac{1}{\epsilon_F\tau}$, an order in which the conductivity is given by the flux-independent Drude formula. The diagrammatic calculation presented below shows that this expectation is indeed correct.

The diagrammatic calculation of Eq. (20) starts by writing

$$\delta(E-H_0)=\frac{1}{2\pi}\left(\frac{1}{E-H_0-i\epsilon}-\frac{1}{E-H_0+i\epsilon}\right)=\frac{1}{2\pi}\left(G^R(E)-G^A(E)\right) \qquad (22)$$

with ϵ being an infinitesimal real number, and $G^R(E), G^A(E)$ being the advanced and retartded Green's functions. Note that H_0 is the adiabatic Hamiltonian, ***including the impurity scattering.*** Employing the conventional impurity technique, we first calculate the contribution of the "classical", Drude–type diagrams (see Fig. 2). Those diagrams are calculated by approximating the Green's function as diagonal in momentum space, with an imaginary part $\frac{1}{\tau_{el}}$ added to the energy (τ_{el} being the elastic mean free time), namely, $G_E^{R\uparrow}(p,p')=\frac{\delta(p-p')}{E-E_p^{\uparrow}+i\frac{1}{2\tau_{el}}}$ and correspondingly for $G^{A\uparrow}, G^{R\downarrow}, G^{A\downarrow}$. The energy $E_P^{\uparrow(\downarrow)}$ is given here by $E_P=\frac{1}{2M}(p-\frac{eB_z\pi a}{2c}\pm\frac{\hbar}{2\pi a}\Omega_g)^2\pm\mu B$. Substituting these Green's functions in Eqs. (20) and (22), and taking only terms of order $\epsilon_F\tau_{el}$, we indeed find that the inverse lifetime is proportional to the Drude expression for the ac conductivity.

$$\frac{1}{\tau}=\frac{D}{(2\pi a)^2}\frac{\pi^2}{2}\frac{\sin^2\alpha}{(2\mu B\tau_{el})^2+1}=\frac{D}{(2\pi a)^2}\frac{\pi^2}{2}\sin^2\alpha\frac{\sigma_{ac}(2\mu B)}{\sigma_{dc}} \qquad (23)$$

where D is the diffusion constant, and σ_{ac}, σ_{dc} are the Drude expressions for the ac and dc real conductivities. Eq. (23) is our first approximation for the impurity averaged lifetime. Before proceeding to improve it, we first use it to get a first approximation for the condition for adiabaticity.

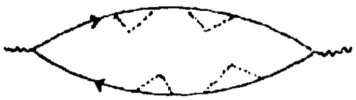


Figure 2: The Drude–type diagrams summed in the expression for the average lifetime, Eq. (23).

For an electron to be non-locally affected by the geometric flux, its spin has to follow the direction of the magnetic field a time long enough such that the geometric phase it accumulates is significant. Hence, when the angle α is of order unity, the lifetime of the adiabatic states, given in Eq. (21), has to be longer than the typical time it takes a diffusing electron to encircle the ring, $\frac{(2\pi a)^2}{D}$. This condition is fulfilled when

$$2\mu B\tau_{el}\gg 1 \qquad (24)$$

Therefore, in the diffusive regime, the adiabatic approximation is valid when the spin precession time is much shorter than the time between elastic scattering events.

In terms of the *ac* conductivity, the adiabatic assumption is valid when the Zeeman energy is large enough so that the *ac* conductivity at the corresponding frequency is much smaller than the *dc* conductivity. Eq. (24) is the condition for adiabaticity also when $\alpha \ll 1$. In that case, the lifetime of an electron has to be long enough for the electron to encircle the ring α^{-1} times – until it accumulates a significant geometric phase. Thus, the condition for adiabaticity becomes $\frac{1}{\tau} \gg \frac{\alpha^2 D}{a^2}$, which reduces to Eq. (24).

The inverse life–time $\frac{1}{\tau}$ was calculated above only to the leading order $\epsilon_F\tau_{el}$. The next order contribution, independent of $\epsilon_F\tau_{el}$, should be calculated by summing the maximally crossed diagrams, the Diffuson and the Cooperon. Again, these diagrams are proportional to those appearing in the calculation of the quantum correction to the conductivity at frequency $2\mu B$, with a difference in the flux affecting each of two Green's functions. However, as long as the Diffuson and the Cooperon are expected to be a small correction to the classical Drude result (that is, as long as $k_F l \gg 1$), Eq. (24) can be accepted as a first approximation to the adiabaticity condition. Then, the Zeeman frequency $2\mu B$ should be of the order of the inverse elastic mean free time. For such a high frequency, the quantum correction to the conductivity is vanishingly small.(8)(9) Therefore, for rings in the metallic regime, where $k_F l \gg 1$, Eq. (24) is the condition for adiabaticity.

We conclude this section by making a few comments regarding the adiabatic condition (24) . First, we interpret its physical origin. As argued by Thouless,(14)contrary to the plane waves eigenstates of free electrons, the single electron eigenstates in a disordered system are superposition of plane waves, with typical spread of $\frac{\hbar}{l}$, where l is the elastic mean free path. In kinetic energy terms, this width is translated into $\frac{\hbar}{\tau_{el}}$. Therefore, the matrix elements of the generalized momentum operator, Π, between states whose kinetic energy differ by more than $\frac{\hbar}{\tau_{el}}$ are negligible. On the other hand, flips of the spin due to H_1 occur only at the Fermi level, i.e., between states whose kinetic energy differ by $2\mu B$. Hence, when the condition (24) is valid, the non adiabatic matrix elements between states at the Fermi level are negligible, and the life–time becomes long. In fact, the condition (24) can be understood also when one considers an electron moving along a typical one dimensional diffusive path $\phi(t)$ (ϕ is again the azimuthal angle describing the electron's position). In the limit of a strong magnetic field, the amplitude of a non–adiabatic spin–flip of the electron is given by(15)(16)

$$a(t) = \int \langle\uparrow(t)|\frac{\partial}{\partial t}|\downarrow(t)\rangle e^{i2\mu Bt} dt \qquad (25)$$

The states $|\uparrow\rangle, |\downarrow\rangle$ depend on time only through the time dependence of the path $\phi(t)$. Thus, the time derivative makes the scalar product $\langle\uparrow(t)|\frac{\partial}{\partial t}|\downarrow(t)\rangle$ proportional to the electron's velocity. The amplitude $a(t)$ becomes exponentially small when the phase of $e^{i2\mu Bt}$ oscillates many times during the characteristic period in which the scalar prduct $\langle\uparrow(t)|\frac{\partial}{\partial t}|\downarrow(t)\rangle$ significantlly varies. This time is the characteristic time during which the velocity varies significantlly, namely, the elastic mean free time. Therefore, when the Zeeman frequency $2\mu B$ is much larger than the inverse elastic mean free time Eq. (25) yields an exponentially small amplitude. Second, we comment that under the strong magnetic fields required to satisfy the condition (24) , one should distinguish between the diffusive limit $\omega_c\tau_{el} \ll 1$ (where ω_c is the cyclotron frequency) and the Landau levels limit $\omega_c\tau_{el} \gg 1$. The relevant limit is determined by the value of the electron's g–factor. Here we assume that the diffusive limit applies. Third, we comment on the relevance of $\frac{1}{\tau}$ to interference effects. As discussed above, the geometric phase accumulated by the electron depends on the direction of its spin. If that direction is flipped at various points along the path, this phase is randomized. Hence, non–adiabatic spin–flips dephase the interference. In the present work we neglect all other mechanisms of dephasing, and therfore τ is to be identified with the phase breaking time τ_ϕ. It is then useful to calculate the ratio of the circumference of the ring to the phase breaking length $L_\phi \equiv \sqrt{D\tau_\phi}$, denoted by Γ

$$\Gamma \equiv \frac{2\pi a}{L_\phi} = \frac{\sqrt{2}\pi \sin\alpha}{\sqrt{1+(2\mu B\tau_{el})^2}} \tag{26}$$

We emphasize that as long as no other dephasing mechanisms are present, this ratio depends neither on the radius a, nor on the temperature T. And finally, we note that for an elastic mean free time of 10^{-11} sec and a g–factor of 10, the adiabaticity condition (24) is satisfied for fields larger than 0.1 Tesla. The ring can be approximated as one dimensional as long as its cross sectional area s satisfies $B_\phi s \ll \phi_0$ (where ϕ_0 is the flux quantum), i.e., as long as it is almost not threaded by magnetic flux created by B_ϕ. For $B_\phi = 0.1$ Tesla, the cross sectional area has to be smaller than $(2000A)^2$.

6. Remnants of the geometric flux in the non–adiabatic case

Our analysis of the effect of the geometric flux on transport properties of the ring has so far concentrated on the adiabatic limit. We now turn to discuss the non–adiabatic limit. Again, we distinguish between ballistic and diffusive rings.

The exact solution of the balistic case was given above, in Eqs. (14)–(16) of section (5). For the convenience of the reader we rewrite the solutions here,

$$|n,\uparrow(\phi)\rangle = e^{in\phi}\begin{pmatrix} \cos\frac{\gamma}{2}e^{-i\phi} \\ -\sin\frac{\gamma}{2} \end{pmatrix} \qquad \text{and} \qquad |n,\downarrow(\phi)\rangle = e^{in\phi}\begin{pmatrix} i\sin\frac{\gamma}{2}e^{-i\phi} \\ \cos\frac{\gamma}{2} \end{pmatrix} \tag{27}$$

The angle γ is implicitly given by

$$\cot\gamma = \cot\alpha + \frac{\hbar^2(2n'-1)}{4Ma^2\mu B\sin\alpha} \tag{28}$$

so that for any finite value of B it is smaller than α. The corresponding eigenvalues are

$$E(n) = \frac{\hbar^2 n^2}{2Ma^2} - \frac{\hbar^2(2n'-1)}{4Ma^2}(1 \pm \cos\gamma) \mp \mu B\cos(\gamma-\alpha) \tag{29}$$

The significance of the angle γ is understood via the calculation of the expectation values of the projection of the spin onto several axes. We calculate these expectation values for the $|n,\uparrow\rangle$ state. The generalization for the $|n,\downarrow\rangle$ states is obvious.

First, we note that the expectation value of σ_z is $\cos\gamma$, i.e., γ is the angle to which the spin bends relative to the z-axis. It is then not surprising to find that the expectation value of the spin projection onto the direction of the magnetic field is $\cos(\gamma-\alpha)$. Two other spin projections of interest are the projection onto two directions perpendicular to the magnetic field, the direction of $\frac{\partial \vec{B}}{\partial \phi}$, which is here the radial direction, and that of $\vec{B}\times\frac{\partial \vec{B}}{\partial \phi}$. It is a matter of simple algebra to find that the former is zero, while the latter is $\sin(\gamma-\alpha)$. The significance of the last two results and their relevance for the understanding of the forces acting on the electron are discussed in Ref. [13].

As seen from the exact solutions Eqs. (27)-(29), when the magnetic field is not strong enough to force the spin to bend in an angle α, the spin bends to a smaller angle $\gamma<\alpha$. The Zeeman energy is then proportional to the projection of the spin onto the magnetic field, and the induced vector potential is still of the form found in the adiabatic case, *but with the angle α replaced by γ*. However, in the adiabatic limit the vector potential was determined only by α and the direction of the spin. Thus, it deserved the name "geometric". In the non-adiabatic case the vector potential depends, through the angle γ, on the magnitude of the magnetic field and the velocity of the electron. Eigenstates of different velocities are then subject to different vector potentials. The vector potential is no more purely geometric.

The observations discussed above in the context of the ballistic case allow for a qualitative understanding of the non-adiabatic limit of the diffusive case. Diffusive eigenstates are built out of superposition of many momentum (or velocity) components. If the magnetic field is too weak to force adiabaticity, each of these components is subject to a different vector potential, and thus also to a different flux. If the range of fluxes induced in the different momentum components is of the order of a flux quantum, the energy of the diffusive eigenstate loses its sensitivity to the direction of the magnetic field, and the geometric effects are lost.

7. How is the geometric flux related to spin-orbit coupling?

Some of the phenomena discussed in this paper, and in particular the multiplicative factor in Eq. (11) are similar to the phenomena that has been shown by Meir, Gefen and Entin-Wohlman[17]to result from a one-dimensional ring of spin-orbit scatterers. It is instructive, then, to devote this section to the relation between the geometric phase and the spin-orbit coupling. This relation becomes clear when the spin-orbit coupling is expressed as a vector potential. The origin of the spin-orbit coupling lies in the coupling of a moving magnetic moment $\vec{\mu}=\frac{e\hbar}{2mc}\vec{\sigma}$ to an electric field $\vec{E}$. In the frame of reference in which the magnetic moment is at rest the electric field is Lorenz-transformed to a magnetic field. If the velocity of the magnetic moment is slow compared to the speed of light, the magnetic field in the rest frame is given by $\frac{\vec{v}}{c}\times\vec{E}$. The magnetic moment couples to that magnetic field in the Zeeman interaction, thus yielding an interaction term $\vec{\mu}\cdot\frac{\vec{v}}{c}\times\vec{E}=\vec{v}\cdot\frac{\vec{\mu}}{c}\times\vec{E}$. Having in mind the interaction term of an electron with an electromagnetic vector potential $\vec{v}\cdot\vec{A}$, we find that $\frac{\vec{\mu}}{c}\times\vec{E} = \frac{e\hbar}{2mc^2}\vec{\sigma}\times\vec{E}$ can be identified as the spin–orbit vector potential. When the magnetic moment arises from the internal spin of a *charged* particle, as in the case of an electron, the acceleration of the particle due to the interaction of the charge with the electric field has to be taken into account, and this leads to a correction factor of $\frac{1}{2}$ to the above expressions. This factor of $\frac{1}{2}$ is known as the Thomas precession factor.[18]Similar to the geometric vector potential discussed in this paper, the spin–orbit vector potential is, in principle, space and spin–dependent, and its values at different points in space do not necessarily commute. It is important, however, to note the differences between the vector potential resulting from the spin–orbit coupling and the one resulting from the Zeeman interaction with a space dependent magnetic field. The first difference has to do with the symmetry with respect to time reversal. While the spin–orbit interaction gives rise to a vector potential, it does not break time–reversal symmetry – it does not induce a $\mp\mu B$ term. Thus, for each eigenstate for which the effective spin–orbit flux is Φ, there is another state, degenerate in energy, for which the effective flux is $-\Phi$. This is Kramers' degeneracy. On the contrary, the effective flux induced by the space–dependent magnetic field is accompanied by the Zeeman energy, that removes the degeneracy. The second difference is a difference in magnitudes. Being inversly proportional to mc^2, the spin–orbit interaction term is very small, unless it invloves very strong electric fields. In the context of condensed matter physics such fields are not "man–made", but rather result from microscopic molecular charge distributions. The microscopic molecular fields are strong enough to make the spin–orbit coupling significant. However, they also vary strongly over microscopic length scales. Thus, when the spin–orbit vector potential results from such microscopic fields, it is a random quantity with a microscopic correlation length. As such, it is uncontrollable, and usually its effect has to be averaged. This averaging gives rise to the weak anti–localization effect.[19]The geometric flux resulting from Berry's phase, on the other hand, is determined by the externally controlable magnetic field. It is also worth noting that while both effects are geometric, i.e., can be expressed as resulting from a vector potential, the origin of their geometric nature is completely different.

Finally, we note that the understanding of the spin–orbit coupling as emerging from a vector potential is useful for a simple analysis of the subject of "hidden momentum" that has attracted some attention in the context of the theory of electromagnetism. [20]

Acknowledgements

I am grateful to Yakir Aharonov for being a superb teacher, advisor and collaborator during my undergraduate, graduate and postgraduate years. I hope to enjoy more of his guidance in years to come.

I am indebted to Y. Imry, Y. Oreg, S. Popescu, S. Levit, A. MacDonald O.

Entin, ?.L. Altshuler, L. Glazman and A. Zee for instructive discussions regarding this work.

REFERENCES

1. M. V. Berry, Proc. R. Lond. **A392**,45 (1984).
2. *Geometric Phases in Physics*, ed. A. Shapere and F. Wilczek, World Scientific, Singapore (1989).
3. A. Stern, Physical Review Letters **68**, 1022 (1992).
4. Y. Imry in *Directions in Condensed Matter Physics* eds. G. Grinstein and G. Mazenko World Scientific, Singapore (1986).
5. B.L. Altshuler, A.G. Aronov, D.E. Khmlenitskii and A.I. Larkin in: *Quantum theory of solids*, ed. I.M. Lifschits, MIR publishers, Moscow (1982).
6. D. Loss, P. Goldbart and A.V. Balatsky Phys. Rev. Lett. **65**, 1655(1990).
7. M. Stone Phys. Rev. D **33**, 1191(1986); J. Moody, A. Shapere and F. Wilczek Phys. Rev. Lett. **56**, 893(1986).
8. Y. Aharonov, E. Ben-Reuven, S. Popescu and D. Rohrlich Phys. Rev. Lett. **65**, 3065(1990) .
9. B.L. Altshuler, A.G. Aronov and B.Z. Spivak Pisma Exsp. Teor. Fiz. **33**, 101(1981), JETP Lett. **33**, 94(1981).
10. S. Chakravarty and A. Schmid, Phys. Rep. **140**, 193 (1986).
11. S. Datta and M. McLennan, Rep. Prog. Phys. **53**, 1003 (1990); A. Aronov, U. Sivan, A. Yacoby, private communications..
12. M.V. Berry *The quantum phase, five years later* in Ref. 2.
13. Y. Aharonov and A. Stern, Physical Review Letters **69**, 3593 (1992).
14. D.J. Thouless, Phil. Mag. **32**, 877(1975) .
15. A. Messiah, *Quantum Mechanics*, Chapter 17, J. Wiley & Sons, New York (1962).
16. D. Bohm, *Quantum theory*, Prentice Hall, New Jersey (1951).
17. Y. Meir, Y. Gefen and O. Entin-Wohlman Phys. Rev. Lett. **63**, 798(1989).

18. J. D. Jackson, *Classical electromagnetism*, Chapter 11 J. Wiley & Sons, New-York (1975); G. Baym, *Lectures on Quantum Mechanics*, Chapter 23, W.A. Benjamin Inc. (1969); An interesting recent derivation of the Thomas precession factor was suggested by H. Mathur, to appear in Phys. Rev. Lett..

19. S. Hikami, A.I. Larkin and Y. Nagaoka, Prog. Theo. Phys. **63**, 707 (1980).

20. L. Vaidman, Am. J. Phys. **58**, 978 (1990); W. Shockley and R.P. James, Phys. Rev. Lett. **18**, 876 (1967); S. Coleman and J.H. Van-Vleck, Phys. Rev. **171**, 1370 (1968).

NEW RESULTS IN THE THEORY OF LANDAU-ZENER TRANSITIONS

ROBERT G. LITTLEJOHN
Department of Physics, University of California
Berkeley, California 94720 USA

ABSTRACT

Adiabatic theory predicts the conservation of quantum numbers in processes with a slow time-dependence, or in systems with slow and fast degrees of freedom. When time scales are not infinitely separated, that is, when there is a breakdown of adiabaticity, then there is some transfer of probability from one slow quantum state to another. This transition probability is given by the famous formula of Landau, Zener, and Stückelberg in the case of coupled, one-dimensional Schrödinger equations. This paper presents a generalization of this formula to general coupled Hermitian systems in one dimension. It is shown that the generalization is almost uniquely determined by the necessary invariance of the transition probability under three groups of transformations, namely, scaling transformations, canonical transformations, and Lorentz transformations. The final formula for the transition probability is a simple function of the simplest quantity one can construct which is invariant under all three of these groups.

The topic of this paper grows out of the theory of adiabatic processes and geometric phases in quantum mechanics, so I will begin by recalling some principal results in this area.

Consider a Hamiltonian which is parameterized by certain parameters R which are slow functions of time:

$$H = H(\mathbf{q}, \mathbf{p}, \mathbf{R}(t)). \tag{1}$$

The usual adiabatic theorem of quantum mechanics asserts that the state,

$$|\psi(t)\rangle = e^{i\gamma(t)}|n(t)\rangle \tag{2}$$

is an approximate solution of the time-dependent Schrödinger equation,

$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = H(\mathbf{R}(t))|\psi(t)\rangle, \tag{3}$$

where $|n(t)\rangle$ is an instantaneous eigenstate of the Hamiltonian,

$$H(\mathbf{R}(t))|n(t)\rangle = E_n(\mathbf{R}(t))|n(t)\rangle, \tag{4}$$

and where the phase $\gamma(t)$ is given by

$$\gamma(t) = -\frac{1}{\hbar}\int^t E_n(\mathbf{R}(t'))\,dt' + \int_{\text{path}} \mathbf{A}(\mathbf{R})\cdot d\mathbf{R}. \tag{5}$$

Here the first term is the so called dynamical phase and the second term is Berry's phase.[1] The differential form in the second term is a 1-form in parameter space, given by

$$A = \mathbf{A} \cdot d\mathbf{R} = i\langle n|dn\rangle. \tag{6}$$

Thus Berry's phase is the line integral of the 1-form A along the path or history of the system $\mathbf{R}(t)$ through parameter space, and the accumulated Berry's phase around a closed loop is given by Stokes' theorem in terms of the closed 2-form $B = dA$.

The 2-form B has singularities in parameter space, similar to the singularity in the magnetic field of a monopole at $\mathbf{r} = 0$. If the Hamiltonian in Eq. (1) has no particular symmetry (as we will assume), then these singularities occur on a manifold of codimension 3 in parameter space. This is because the singular manifold is surface on which the energy level $E_n(\mathbf{R})$ is degenerate with another level, $E_n(\mathbf{R}) = E_m(\mathbf{R})$. These singularities serve as sources for Berry's curvature form B.

However, the condition which must be satisfied for the adiabatic theorem to be valid is that energy levels must be well separated. More quantitatively, the condition is

$$\frac{\dot{H}}{H} \ll \frac{E_n - E_m}{\hbar}, \tag{7}$$

which is a way of saying that the transition frequency between the level E_n of interest and the closest other level E_m must be large in comparison to the typical frequency component of the Hamiltonian H. Therefore if the history of the system $\mathbf{R}(t)$ should pass close to the sources of Berry's 2-form on the singularity manifold, then the adiabatic theorem and the results quoted in Eqs. (2)–(6) will break down. Let us therefore introduce a perturbation parameter,

$$\epsilon = \frac{\hbar \dot{H}}{H(E_n - E_m)}, \tag{8}$$

so that adiabatic theory can be systematically developed as an expansion in powers of ϵ. (More precisely, ϵ is a typical value of the right hand side of Eq. (8), or a scaling parameter for a family of systems.) Then we find that the results quoted in Eqs. (2)–(6) above are the leading terms in an expansion in ϵ, and that there are higher order terms which can be worked out. For example, Berry's phase is a correction which is of order ϵ in comparison to the dynamical phase.

Now let us generalize the situation, and allow the parameters to become dynamical variables themselves. That is, let us replace $\mathbf{R}$ by $(\mathbf{Q}, \mathbf{P})$, which are slow degrees of freedom, so that the (now time-independent) Hamiltonian reads,

$$H = H(\mathbf{q}, \mathbf{p}; \mathbf{Q}, \mathbf{P}), \tag{9}$$

where $(\mathbf{q}, \mathbf{p})$ are the fast degrees of freedom as before. The best known example of a Hamiltonian of this type is the Born-Oppenheimer Hamiltonian which is so useful in molecular physics. We may allow the slow degrees of freedom to be either classical or quantum mechanical, but, even in the case in which they are quantum mechanical variables, it is often useful to treat them by semiclassical methods. This is because the separation of time scales often implies that the slow quantum numbers are large (as when slow and fast energies are comparable).

Therefore in either case it is appropriate to think of a classical phase space for the slow degrees of freedom, which becomes identified with the parameter space discussed above. This classical $(\mathbf{Q},\mathbf{P})$ phase space naturally supports the symplectic 1-form $\theta_S = \mathbf{P}\cdot d\mathbf{Q}$ as do all classical phase spaces, but it also supports the 1-form for Berry's phase, $\theta_B = i\langle n|dn\rangle$. It is geometrically reasonable that these two 1-forms should be linked somehow, and, indeed, as shown first by Kuratsuji and Iida,[2] there is an effective symplectic 1-form which is the sum of the two,

$$\theta_{\text{eff}} = \mathbf{P}\cdot d\mathbf{Q} + i\hbar\langle n|dn\rangle, \tag{10}$$

which governs the semiclassical quantization of the slow degrees of freedom. That is, when the slow degrees of freedom are viewed on a semiclassical level, the average effect of the fast degrees of freedom appear as a modification of the classical symplectic form. Greg Flynn and I have developed these issues in the context of WKB theory, and explored some examples.[3]

We now introduce some fixed basis $|\alpha\rangle$ for the fast degrees of freedom. By "fixed" we mean that these basis vectors do not depend on the slow variables $(\mathbf{Q},\mathbf{P})$; for example, in the Hamiltonian for a molecule, we could introduce a harmonic oscillator basis for the electronic wave functions. Then the Hamiltonian of Eq. (9) becomes a matrix in the fast indices,

$$H(\mathbf{q},\mathbf{p};\mathbf{Q},\mathbf{P}) \rightarrow H_{\alpha\beta}(\mathbf{Q},\mathbf{P}), \tag{11}$$

and the Schrödinger equation becomes a system of coupled wave equations in the slow variables:

$$\left[H_{\alpha\beta}(\mathbf{Q},\mathbf{P}) - E\,\delta_{\alpha\beta}\right]\psi_\beta(\mathbf{Q}) = 0. \tag{12}$$

For example, the molecular Hamiltonian has the Born-Oppenheimer form,

$$\left[\left(\frac{P^2}{2M} - E\right)\delta_{\alpha\beta} + V_{\alpha\beta}(Q)\right]\psi_\beta(Q) = 0, \tag{13}$$

where $V_{\alpha\beta}$ is a matrix of potential energies. There are no gauge terms in Eq. (13) because we have used a fixed basis. More generally, we have a system of coupled wave equations which we write in the form,

$$D_{\alpha\beta}(\mathbf{Q},\mathbf{P})\,\psi_\beta(\mathbf{Q}) = 0, \tag{14}$$

where D is a matrix of operators in the slow variables. It is Eq. (14) which we wish to treat by semiclassical methods, making as few assumptions as possible about the operators which appear as the components of D.

As is well known, semiclassical wave functions are represented in the classical phase space by means of so-called Lagrangian manifolds,[4] which are N-dimensional surfaces in the $2N$-dimensional phase space upon which the symplectic 2-form vanishes. Here N is identified with the number of slow degrees of freedom. As long as adiabatic conditions are satisfied, the WKB solutions of Eq. (14) can be developed in manner which is much like standard semiclassical theory, except for interesting issues regarding the gauge form θ_B and its role in quantization. I will not go into this here, but rather I will devote the rest of this paper to another question, namely, what happens if the Lagrangian manifold of dimensionality N should pass close to the singularity manifold of codimension 3? This latter manifold can be seen as the manifold upon which the matrix D, regarded as a function of classical variables $(\mathbf{Q},\mathbf{P})$, has a double vanishing eigenvalue, i.e., it has a corank of 2 or more.

The answer, roughly speaking, is that there will be nonadiabatic transitions between the fast eigenstates $|n\rangle$ and $|m\rangle$. These are the so-called Landau-Zener-Stückelberg transitions, and the process is sometimes called "mode conversion." The original treatment of Landau,[5] Zener,[6] and Stückelberg[7] was applied to the case of coupled Schrödinger equations of the form of Eq. (13) in one slow degree of freedom. They derived the transition probability,[8]

$$T = \exp\left(-\frac{2\pi\Delta^2}{\hbar v_0|V'_{11} - V'_{22}|}\right), \tag{15}$$

where Δ, V_{11}, and V_{22} are parameters of the potential energy matrix at the mode conversion point, where the prime indicates an $X = Q$ derivative, and where v_0 is the velocity at which the particle moves through the mode conversion region. This case has been subject to sixty years of investigation, and is now quite well understood. For our purposes, the important thing to notice about this result is that it scales as $e^{-1/\epsilon}$ in the adiabatic perturbation parameter introduced in Eq. (8). Thus, we see that these nonadiabatic transition probabilities are beyond all orders in ϵ and cannot be obtained by straightforward perturbation methods.

Coupled Schrödinger equations in higher numbers of slow degrees of freedom are important in molecular scattering theory, and are still an active area of research. For more general wave equations of the type shown in Eq. (14), special cases have been studied in one slow degree of freedom, but almost nothing is known about the case of higher degrees of freedom. For the rest of this paper I will concentrate on the case of mode conversion in one slow degree of freedom, treating the general case indicated in Eq. (14). I will henceforth write (Q,P) for the slow variables (in italic type), since there is only one degree of freedom.

Thus we consider coupled wave equations of the form,

$$D_{\alpha\beta}(Q,P)\,\psi_\beta = 0. \tag{16}$$

The matrix D of slow operators can be of any size, but without essential loss of generality it can be restricted to a 2×2 matrix. This is because the breakdown of adiabaticity, when it occurs, generically only involves two different levels E_n and E_m. Of course it is possible that more could be involved, and there is the very interesting possibility of global degeneracies, but here for simplicity we will take the most generic case which is that of two interacting levels. Then one can show that adiabatic transformations can be used to reduce the original system to a 2×2 system, essentially by block diagonalizing the original D matrix and leaving a 2×2 block on the diagonal.

Accepting this, we can write the coupled wave equations in the form,

$$\begin{pmatrix} D_{11}(Q,P) & D_{12}(Q,P) \\ D_{12}(Q,P)^{\dagger} & D_{22}(Q,P) \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = 0. \tag{17}$$

Since we sometimes think of the slow variables (Q,P) in a semiclassical sense, we will sometimes treat the matrix D as a matrix of classical functions of (Q,P) (not operators). Thus, D becomes a Hermitian matrix field over the slow phase space. There will be a breakdown of adiabaticity and subsequent Landau-Zener-Stückelberg transitions between fast energy levels when both eigenvalues of this matrix are small in some region of phase space. Our goal will be to compute the transition probability T in such a case, and thereby generalize the Landau-Zener-Stückelberg formula given in Eq. (15).

We will base this computation on symmetry arguments. We argue that the transition probability T must be a function of $D_{\alpha\beta}$ and its derivatives with respect to Q and P which is invariant under three classes of symmetry operations. These transformations are scaling transformations, symplectic or canonical transformations, and Lorentz transformations. We will now explain these transformations in greater detail.

The scaling transformations involve simply multiplying Eq. (17) through by some constant a, so that $D \to aD$. Such a transformation of course changes nothing essential about the wave equation itself, and the transmission probability T must therefore be invariant, $T \to T$. This implies that T must be a homogeneous function of degree 0 of $D_{\alpha\beta}$ and its derivatives.

Next we invoke canonical or symplectic invariance. It is now well understood that when quantum mechanical quantities which are independent of representation, such as energy levels or transition probabilities, are computed by semiclassical means, then the semiclassical expression must be a canonical invariant. A nice example of this is the Bohr-Sommerfeld or EBK formula for energy levels; the energy levels are given in terms of classical actions, which are invariant under canonical transformations. In the present case, we expect T to be invariant under canonical transformations, which means that all Q and P derivatives of $D_{\alpha\beta}$ which occur in the expression for T must be expressible in terms of Poisson brackets.

The third class of transformations involves Lorentz invariance. If we replace the 2-component ψ-field shown in Eq. (17) by a constant linear transformation of itself,

$$\psi = Q\psi', \tag{18}$$

where Q is any invertible 2×2 matrix (possibly complex), then the Hermiticity of the equations is preserved if we write

$$D' = Q^{\dagger} D Q. \tag{19}$$

Obviously the transition probability cannot change under such a transformation, so we expect the formula for T in terms of $D_{\alpha\beta}$ to be invariant when D is replaced by D' as in Eq. (19). All we require of the matrix Q is that it be invertible, but, without loss of generality, Q can be restricted to have unit determinant, since if the determinant is not unity, it can be made so by a scaling transformation such as those we have already considered. Thus, Q can be restricted to the group $SL(2,\mathbb{C})$, the spinor representation of the Lorentz group.

One might have thought that unitary transformations would be sufficient to solve the problem at hand, but this turns out not to be the case; in order to obtain the necessary normal forms which underlie this generalized Landau-Zener-Stückelberg theory, it is necessary to invoke nonunitary transformations.

To bring out the Lorentz invariance more clearly, we write

$$D(Q,P) = B^{\mu}(Q,P)\,\sigma_{\mu}, \tag{20}$$

where $\sigma_{\mu} = (I,\sigma_x,\sigma_y,\sigma_z)$ is the usual 4-vector of Pauli matrices, so that B^{μ} is a 4-vector field defined over the slow phase space. Then under the transformation of Eq. (19), the 4-vector B^{μ} transforms according to

$$B^{\mu} \to \Lambda^{\mu}{}_{\nu} B^{\nu}, \tag{21}$$

where $\Lambda^{\mu}{}_{\nu}$ is a 4×4 Lorentz transformation. Therefore the transition probability T must be a Lorentz scalar when expressed in terms of the 4-vector B^{μ}.

Altogether, we require a quantity which is a simultaneous Lorentz scalar and a symplectic scalar, and a homogeneous function of B^{μ} of degree 0. We begin by listing the simplest simultaneous Lorentz and symplectic scalars we can write down. We use Poisson brackets (denoted by curly brackets) to guarantee that we have a symplectic scalar. The simplest four such scalars are the following:

$$B^{\mu}B_{\mu} = \det D, \tag{22a}$$

$$\{B^{\mu},B_{\mu}\} = 0, \tag{22b}$$

$$B^{\mu}B^{\nu}\{B_{\mu},B_{\nu}\} = 0, \tag{22c}$$

$$\{B^{\mu},B^{\nu}\}\{B_{\mu},B_{\nu}\} \neq 0. \tag{22d}$$

Of these, the middle two vanish identically because of the antisymmetry of the Poisson bracket and the symmetry of the Lorentz contraction. The first and the fourth are the simplest nonvanishing scalars with the required invariance properties, of these, the first is a homogeneous function of B^{μ} of degree 2, and the fourth is a homogeneous function of degree 4. Therefore the simplest homogeneous function of degree 0 we can create with the required invariance properties is obtained by dividing the first scalar by the square root of the fourth scalar. We expect that the Landau-Zener transition probability T must be a function of this quantity.

Indeed, a more detailed calculation gives the result in the form,

$$T = \exp\left(-\frac{2\pi}{\hbar}\frac{B^{\mu}B_{\mu}}{\sqrt{2\{B^{\mu},B^{\nu}\}\{B_{\mu},B_{\nu}\}}}\right). \tag{23}$$

This is the required generalization of the Landau-Zener formula, and is our principal result for this paper.

One might find this "derivation" somewhat unsatisfying, in that the invariance principles alone do not allow us to determine the final functional form of the transition probability, as displayed in Eq. (23). But the more detailed calculation just alluded to involves using the three transformation groups we have discussed to transform the original coupled wave equation in Eq. (17) into a standard or normal form, which is then solved by standard analytic methods. The invariance properties of these transformation groups are an important aspect of the normal form transformations. Thus it is not misleading to emphasize the importance of symmetry principles in discussing the derivation of Eq. (23).

The transformation groups we have discussed here are also important in the treatment of Landau-Zener transitions in many dimensions, including the case of multidimensional Born-Oppenheimer problems. We will report on such calculations in the future.

Acknowledgements

This work was supported by the U. S. Department of Energy under contract Nos. DE-AC03-76SF00098 and W-7405-Eng-48, and by the Plasma Physics Research Institute of Lawrence Livermore National Laboratory.

References

1. Alfred Shapere and Frank Wilczek, *Geometric Phases in Physics* (World Scientific, Singapore, 1989).
2. H. Kuratsuji and S. Iida, *Prog. Theor. Phys.* **74** (1985) 439; *Phys. Lett.* **A111** (1985) 220; *Phys. Lett.* **B184** (1987) 242; *Phys. Rev.* **D37** (1988) 441.
3. R. G. Littlejohn and William G. Flynn, *Phys. Rev.* **A44** (1991) 5239; *Chaos* **2** (1992) 149; *Phys. Rev.* **A45** (1992) 7697; "General linear mode conversion coefficient in one dimension," in press, Phys. Rev. Lett., 1993.
4. V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
5. L. D. Landau, *Phys. Z. Sowietunion* **1** (1932) 88.
6. Clarence Zener, *Proc. Roy. Soc.* **A137** (1932) 696.
7. E. C. G. Stückelberg, *Helv. Phys. Acta* **5** (1932) 369.
8. E. E. Nikitin and S. Ya. Umanskii, *Theory of Slow Atomic Collisions* (Springer-Verlag, New York, 1984).

QUANTUM MECHANICS OF THE ELECTRIC CHARGE

A. STARUSZKIEWICZ
Institute of Physics, Jagellonian University, Reymonta 4
30 059 Kraków, Poland

ABSTRACT

A simple argument against the existence of magnetic monopoles is given. The argument is an important part of the quantum theory of the electric charge developed by the author.

> "The same modification of the (Maxwell - Lorentz) theory which contains e as a consequence, will also have the quantum structure of radiation as a consequence."
>
> *Albert Einstein*
> (*Phys. Zeit.* 10 (1909) 192)

1.Introduction

This paper is dedicated to Professor Yakir Aharonov on the occasion of his 60th birthday. The subject of the paper, quantum mechanics of the electric charge, is based on the notion of *phase*, this elusive concept which has always fascinated Professor Aharonov.

The electric charge Q and the phase $S(x)$ of a (second quantized) charged system are canonically conjugated variables:

$$[Q, S(x)] = ie, \quad (\hbar = 1 = c) \tag{1}$$

e being the elementary charge. Proof of this theorem is given in [1]. Here I will make only two rather obvious comments.

Eq.(1) does explain quantization of the electric charge Q in units equal to the constant e:

$$Q = ne, \quad n = 0, \pm 1, \pm 2, \ldots$$

It does not, however, explain the universality of the electric charge i.e. the fact that e.g. the electric charge of the electron seems to be mathematically equal to the electric charge of the proton. Indeed, since the constant e in Eq.(1) is arbitrary, we cannot exclude theoretically a situation in which $e = e_1$ for one charged system and $e = e_2 \neq e_1$ for another system.

2. The phase $S(x)$ can be uniquely determined at the spatial infinity

x in Eq.(1) is an arbitrary spatio-temporal point. Let us imagine that x

tends to the spatial infinity:

$$xx \equiv (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 \to -\infty.$$

Mathematically-minded readers will object that we are not allowed to fix, even in the form of a limit, the argument of an operator-valued distribution. True. The argument which follows is physical rather than mathematical, it constitutes a piece of theoretical rather than mathematical physics.

At the spatial infinity there is only one function which can possibly play the role of phase. This function must be equal to

$$S(x) = -ex^\mu A_\mu(x), \tag{2}$$

where e is a constant proportionality factor and $A_\mu(x)$ is the electromagnetic potential. To see this one has to note that at the spatial infinity the electromagnetic field is free,

$$\partial^\mu F_{\mu\nu} \equiv 4\pi j_\nu = 0$$

and homogeneous of degree -2, $F_{\mu\nu}(\lambda x) = \lambda^{-2} F_{\mu\nu}(x)$ for each $\lambda > 0$ [2]. The field is free because the electric current j_ν, being carried by massive particles, must be confined to the future and past light cone. It must be homogeneous of degree -2 because, as seen e.g. in the static case, the charge generated monopole term dominates dipole and higher terms.

Consider a classical electromagnetic field which is free and homogeneous of degree -2; assume that its potential is homogeneous of degree -1, which is natural. Let us form two vectors,

$$F_{\mu\nu}(x)\,x^\nu \quad \text{and} \quad \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} x_\nu F_{\rho\sigma}(x),$$

where x is the radius vector in the Lorentzian reference frame in which the homogeneity condition holds.

The two vectors given above determine the tensor $F_{\mu\nu}$ in a purely algebraic way. Both these vectors are gradients of homogeneous of degree zero functions:

$$F_{\mu\nu}(x)\,x^\nu = \partial_\mu e(x), \quad \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} x_\nu F_{\rho\sigma}(x) = \partial^\mu m(x).$$

$e(x)$ and $m(x)$ denote "electric" and "magnetic" parts respectively. $e(x)$ can be easily calculated:

$$\begin{aligned} F_{\mu\nu}(x)\,x^\nu &= [\partial_\mu A_\nu(x) - \partial_\nu A_\mu(x)]\,x^\nu = \\ &\partial_\mu [A_\nu(x)\,x^\nu] - \delta^\nu_\mu A_\nu(x) - x^\nu \partial_\nu A_\mu(x) = \partial_\mu [x^\nu A_\nu(x)] \end{aligned}$$

because

$$x^\nu \partial_\nu A_\mu(x) = -A_\mu(x)$$

from the Euler theorem on homogeneous functions.

I maintain that $m(x)$ must be a constant. This is an argument against the existence of magnetic monopoles which, to the best of my knowledge, has never been put forward before. (The argument given by Dr. Herdegen [3] is different.)

To see this let us calculate the Lagrangian density

$$dx^0dx^1dx^2dx^3F_{\mu\nu}F^{\mu\nu} \tag{3}$$

for a homogeneous of degree -2 field $F_{\mu\nu}$, using the spherical coordinates

$$\begin{aligned} x^0 &= \xi^0\sinh\xi^1, \\ x^1 &= \xi^0\cosh\xi^1\sin\xi^2\cos\xi^3, \\ x^2 &= \xi^0\cosh\xi^1\sin\xi^2\sin\xi^3, \\ x^3 &= \xi^0\cosh\xi^1\cos\xi^2, \end{aligned}$$

$$0<\xi^0<\infty, \quad -\infty<\xi^1<+\infty, \quad 0\leq\xi^2\leq\pi, \quad 0\leq\xi^3<2\pi.$$

These coordinates cover in an obvious way the spatial infinity we are interested in. Note that ξ^0 is a space-like coordinate while ξ^1 is a time-like coordinate. A simple calculation gives

$$dx^0dx^1dx^2dx^3F_{\mu\nu}F^{\mu\nu} = 2\frac{d\xi^0}{\xi^0}\sqrt{g}\,d\xi^1d\xi^2d\xi^3\left(-g^{ik}\partial_i e\,\partial_k e + g^{ik}\partial_i m\,\partial_k m\right).$$

Here

$$g_{ik} = (\xi^0)^{-2}g_{\mu\nu}\frac{\partial x^\mu}{\partial\xi^i}\frac{\partial x^\nu}{\partial\xi^k}, \quad i,k=1,2,3,$$

is the metric on the spatial infinity.

The Lagrangian density (3) is seen to be a difference of two identical Lagrangian densities. Thus only one of them can have the correct sign i.e. the sign which, upon quantization, would give a positive definite inner product. The part with the right sign is *called* electric, the part with the wrong sign is *called* magnetic and must be put equal to zero.

Now, the Gauss theorem says that the total charge Q is determined by the electromagnetic field at the spatial infinity. In the quantum theory the charge operator Q must have its canonically conjugated variable $S(x)$. Thus $S(x)$ must have a "tail" which does not vanish even at the spatial infinity. We have seen, however, that there is exactly one function, namely $x^\mu A_\mu(x)$, which can play the role of the "tail". Hence, there must exist a constant e such that at the spatial infinity

$$S(x) = -ex^\mu A_\mu(x). \tag{2}$$

The constant e in this equation is identical with the constant e in Eq.(1). This is a hypothesis substantiated in the next section.

3. The proportionality factor in the phase

The two equations

$$\boxed{\begin{aligned}[Q, S(x)] &= ie, \\ S(x) &= -ex^{\mu}A_{\mu}(x),\end{aligned}}$$

constitute together a closed theory, the quantum mechanics of the electric charge. It is important to understand correctly the epistemological status of both equations. The first equation is simply a theorem in the Q.E.D. which, by continuity, is assumed to hold also at the spatial infinity. The second equation is a hypothesis; one can give several arguments supporting Eq.(2) but all those arguments do not amount to a proof. Here are two simple arguments, to be added to those which I have given elsewhere [1].

Take the Coulomb field of the charge Q at rest:

$$A_0 = \frac{Q}{r}, \quad A_1 = A_2 = A_3 = 0.$$

Its phase, according to Eq.(2), is

$$S(x) = -e\frac{Q}{r}t = -eQ\frac{t}{r}.$$

During the eternity of time available at the spatial infinity,

$$-r < t < r,$$

the phase $S(x)$ changes from eQ to $-eQ$. Take now the hydrogen atom with the nuclear charge Q and the electron charge e and assume that the radius of its circular orbit tends to infinity. During the eternity of time available,

$$-r < t < r,$$

the electromagnetic phase of the electron wave function,

$$-e\int A_{\mu}(x)\,dx$$

will change by the same amount:

$$-e\int_{-r}^{r}\frac{Q}{r}\,dt = -2eQ.$$

Thus the phase given by Eq.(2) changes as the true phase of the electron wave function in an infinitely large hydrogen atom.

The phase of the Coulomb field ,

$$S(x) = -\frac{eQ}{r}t$$

may be compared with the phase of the wave function of a stationary state, $-Et$, E being the energy of the stationary state. Thus $S(x)$ looks like the phase of a stationary state driven by the Coulomb energy eQ/r. Again, this is not a proof but a heuristic argument supporting Eq.(2).

Equations (1) and (2) together do allow to explain the universality of the electric charge. To be more precise, they allow to prove the fol!owing theorem: the total charge of the universe is always a multiple of a single constant. To apply this to the electron or to the proton one must be able to estimate the accuracy with which, under specific observational circumstances, they can be considered as isolated universes. The experimental equality of electron's and proton's charge shows that this accuracy is indeed extremely high.

4. Acknowledgement

I thank Professor Paweł O. Mazur for encouragement and several useful discussions.

5. References

1. A. Staruszkiewicz, *Ann. Phys. (N.Y.)* **190** (1989) 354.
2. J. L. Gervais and D. Zwanziger, *Phys. Lett.* **B94** (1980) 389.
3. A. Herdegen, *J. Phys. A*. **26** (1993) L449.

Magnetic charges and local duality symmetry

Patrick Das Gupta

Department of Physics and Astrophysics,

University of Delhi, Delhi - 110 007

Abstract

The notion of magnetic charge is intimately linked with the global duality symmetry exhibited by the extended Maxwell equations. It is easy to show that duality symmetry is meaningful only in 3+1 dimensional space-times, implying thereby that magnetic monopoles as fundamental particles can be postulated only in 3+1 dimensions. It is interesting to study the consequences of elevating the status of duality symmetry to a local symmetry. This is achieved by introducing a complex scalar field in a theory that treats electric and magnetic charge on equal footing. The new theory is a generalization of the extended Maxwell theory, which reduces to the usual Maxwell electrodynamics in the low energy. The electric charge arises due to a spontaneous symmetry breaking in the scalar field sector. A suitable choice of gauge makes the magnetic charge vanish.

1. Introduction

Magnetic charge as a concept is very interesting because of several reasons. Firstly, they make the structure of classical electrodynamics more symmetric. The second reason is that existence of a single magnetic monopole in the universe can explain charge quantization [1]. Furthermore, 't Hooft [2] and Polyakov [3] showed that magnetic monopoles are generic in grand unified theories. In section 2, we show that the duality symmetry is meaningful only in 3+1 dimensional space-times, implying that the notion of magnetic charge is linked with the dimensionality of space-time. Then, we gauge the duality symmetry by invoking a complex scalar field. Finally, in section 3, we show that the resulting theory is a generalization of standard electrodynamics, which reduces to the usual Maxwell equations when there is a spontaneous symmetry breaking in the scalar field sector. In our model, although we start off with a theory in which electric and magnetic charge have the same rank, we get the interesting result that magnetic charge can be gauged away.

2. Local duality symmetry

In 3 + 1 dimensional flat space-time, when magnetic monopoles are present, electromagnetic theory is described in terms of extended Maxwell-Lorentz equations [4],

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} j_e^\nu, \tag{1}$$

$$\partial_\mu \tilde{F}^{\mu\nu} = \frac{4\pi}{c} j_m^\nu, \tag{2}$$

and,

$$\frac{dp^\mu}{d\tau} = \frac{1}{c}\left[q_e F^{\mu\nu} + q_m \tilde{F}^{\mu\nu}\right]\frac{dx_\nu}{d\tau}, \tag{3}$$

where $\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu}{}_{\alpha\beta}F^{\alpha\beta}$ is the dual of the electromagnetic field tensor $F^{\mu\nu}$, while j_e^μ and j_m^μ are the 4-current densities corresponding to electric and magnetic charges, respectively.

It can be easily verified that under the following transformation,

$$F_{\mu\nu} \to F'_{\mu\nu} = \cos\theta F_{\mu\nu} - \sin\theta \tilde{F}_{\mu\nu}, \tag{4}$$

and,

$$q_e \to q'_e = \cos\theta q_e - \sin\theta q_m, \tag{5}$$

$$q_m \to q'_m = \sin\theta q_e + \cos\theta q_m, \tag{6}$$

extended Maxwell equations (1) - (3) are invariant.

Is the duality rotation (4) - (6) meaningful for electrodynamics in space-times of arbitrary dimensions ? To answer this, we consider electrodynamics without magnetic charges in D+1 dimensional flat space-time. The action corresponding to a particle of charge q interacting with electromagnetic fields is given by,

$$\mathcal{A} = -mc \int \sqrt{\eta_{\mu\nu} dx^\mu dx^\nu} - \frac{q}{c} \int A_\mu dx^\mu - \frac{1}{16\pi} \int F_{\mu\nu} F^{\mu\nu} d^{D+1}x, \tag{7}$$

where $\mu, \nu = 0, 1, 2, ..., D$.

The equations of motion that follow from (7) are,

$$E^i, i = 4\pi\rho, \tag{8}$$

$$F^{ij}, j = -\frac{1}{c}\frac{\partial E^i}{\partial E} - \frac{4\pi}{c} j^i, \tag{9}$$

$$\tilde{F}^{\mu_1\mu_2\cdots\cdots\mu_{D-1}}, \mu_1 = 0, \tag{10}$$

where $E^i \equiv -F^{0i}$ and $\tilde{F}^{\mu_1\mu_2\cdots\cdots\mu_{D-1}} = \frac{1}{2}\epsilon^{\mu_1\mu_2\cdots\mu_{D-1}\mu_D\mu_{D+1}} F_{\mu_D\mu_{D+1}}$ with $i, j = 1, 2, ..., D$ and $\mu, \nu = 0, 1, 2, ..., D$. In order to extend eqs (8) - (10) by adding magnetic monopoles, one needs to modify eq (10) so that $\tilde{F}^{\mu_1\mu_2\cdots\mu_{D-1}}, \mu_1 = \frac{4\pi}{c} j_m^{\mu_2\mu_3\cdots\mu_{D-1}}$. But this immediately brings an asymmetry between electric and magnetic charges, because the electric charge current density is only a $(D+1)$-vector. There is symmetry only when $\tilde{F}^{\mu_1\mu_2\cdots\mu_{D-1}}$ and $F^{\mu_1\mu_2}$ are of same rank, implying $D = 3$. Therefore, we conclude that duality transformation is a meaningful symmetry only when the dimensionality of space-time is $3+1$, implying that only in such space-times electric and magnetic charges have similar status.

Duality symmetry (4) - (6) is a U(1) symmetry. To see this, we define complex electromagnetic field tensor, complex charge and current density, respectively, as

$$G_{\mu\nu} = F_{\mu\nu} + i\tilde{F}_{\mu\nu}, \tag{11}$$

$$Q = q_e + iq_m, \tag{12}$$

and,

$$J^\mu = j_e^\mu + ij_m^\mu, \tag{13}$$

we can re-write the extended Maxwell equations (1) and (2) as,

$$\partial_\mu G^{\mu\nu} = \frac{4\pi}{c} J^\nu, \tag{14}$$

and the generalized Lorentz force equation (3) as,

$$\frac{dp^\mu}{d\tau} = \frac{1}{2c}\left[Q^* G^{\mu\nu} + QG^{*\mu\nu}\right]\frac{dx_\nu}{d\tau}. \tag{15}$$

Because of (4) - (6), the duality rotation now reads as,

$$G_{\mu\nu} \rightarrow G'_{\mu\nu} = e^{i\theta} G_{\mu\nu}, \tag{16}$$

$$Q \rightarrow Q' = e^{i\theta} Q, \tag{17}$$

and,

$$J^{\mu} \rightarrow J'^{\mu} = e^{i\theta} J^{\mu}. \tag{18}$$

It is obvious that the equations (14) and (15) are invariant under the transformation (16) - (18). So far we had been assuming that the transformation parameter θ is constant in space-time, implying that the duality transformation is global. We wish, now, to extend the hitherto global symmetry to a local one, by making θ depend on space-time coordinates. This clearly requires modification of the field equations. More significantly, local duality transformation makes the electromagnetic charge Q space-time dependent! This is an unusual feature suggesting a different way of looking at the concept of electromagnetic charge. In the next paragraph we elaborate on this.

To begin with, we introduce a complex scalar field $\phi(x)$ which under local duality transformation changes as follows,

$$\phi(x) \rightarrow \phi'(x) = e^{i\theta(x)} \phi(x) \quad . \tag{19}$$

In this new picture, the electromagnetic charge arises due to the interaction between the charged particle and the scalar field ϕ so that,

$$Q(x(\tau)) = \alpha\phi(x(\tau)) \quad , \tag{20}$$

where $x^{\mu}(\tau)$ is the world line of the particle and α is a coupling constant that solely depends on the particle. This way of viewing at the electromagnetic charge is reminiscent of the origin of mass in electroweak theories through Higgs field. In fact, in the next section we will incorporate most of the features associated with the Higgs sector in the dynamics of ϕ.

We now make use of the scalar field ϕ to define a gauge covariant derivative,

$$\mathcal{D}_\mu \equiv \partial_\mu - \psi_\mu \tag{21}$$

where,

$$\psi_\mu = \frac{\partial_\mu \phi(x)}{\phi(x)} \quad . \tag{22}$$

From (21) and (22), it is easy to see that,

$$\mathcal{D}_\mu G^{\alpha\beta} \rightarrow e^{i\theta(x)} \mathcal{D}_\mu G^{\alpha\beta}. \tag{23}$$

In (21) ψ_μ acts apparently like a gauge field, but (22) makes it obvious that this is a pure gauge. And, hence, the definition (21) does not introduce any new gauge interaction. Modifying (14) to ,

$$\mathcal{D}_\mu G^{\mu\nu} = \frac{4\pi}{c} J^\nu, \tag{24}$$

we find that the equations of motion given by (15) and (24) are invariant under the local duality transformation, and these form the generalized version of the extended Maxwell equations. In the following section, we will derive these equations as well as the equations of motion for the scalar field from an action.

3. Lagrangian formulation

In this section, we derive the equations of motion for fields and particles from an action. We begin with few definitions. Let $a_\mu(x)$ be a complex 4-vector field that under duality transformation behaves in the following way :

$$a_\mu(x) \rightarrow a'_\mu(x) = e^{i\theta(x)} a_\mu(x). \tag{25}$$

The complex electromagnetic field tensor $G_{\mu\nu}$ is related to a_μ in the following way,

$$G_{\mu\nu} = (\partial_\mu + \psi^*_\mu)a_\nu - (\partial_\nu + \psi^*_\nu)a_\mu \quad , \tag{26}$$

where ψ_μ is related to ϕ according to (22). However, not all the components of a_μ are independent. This is because of the definition (11) for $G_{\mu\nu}$ that requires the following constraint to be satisfied,

$$G_{\mu\nu} = \frac{i}{2}\epsilon_{\mu\nu}{}^{\alpha\beta}G_{\alpha\beta}. \tag{27}$$

The action for the electromagnetic field is given by,

$$\mathcal{A}_0 = -\frac{1}{16\pi}\int G^*_{\mu\nu}G^{\mu\nu}d^4x \quad . \tag{28}$$

For particles, we label the world-lines $y^\mu(\tau)$ with latin indices i,j,.. = 1,2,..., and denote the world-line and the 4-velocity of the j^{th} particle as $y^\mu(\tau_j) \equiv y^\mu_j$ and $\frac{dy^\mu}{d\tau_j} \equiv \dot{y}^\mu_j$. The portion of the total action relevant for the equations of motion corresponding to particles is given by,

$$\mathcal{A}_1 = -\sum_j m_j c\int\sqrt{\eta_{\mu\nu}\dot{y}^\mu_j\dot{y}^\nu_j}d\tau_j \; - \; \frac{1}{2c}\sum_j \alpha_j\int[\phi^*(y_j)a^\mu(y_j) + \; \text{c.c.}]\dot{y}_{\mu j}d\tau_j \quad , \tag{29}$$

where c.c. denotes the complex conjugate, and α_j is the coupling constant (see (20)) orresponding to the j-th particle.

We now come to the scalar field ϕ. Because of (19), the scalar field sector has to be invariant under local U(1) group suggesting the existence of a abelian gauge field χ_μ that interacts with ϕ. The corresponding gauge covariant derivative then can be written as,

$$\nabla_\mu = \partial_\mu - ig\chi_\mu \quad , \tag{30}$$

where g is the gauge coupling constant.

Under local duality transformation, the abelian gauge field transforms as,

$$\chi_\mu \rightarrow \chi'_\mu = \chi_\mu + \frac{1}{g}\partial_\mu\theta \quad . \tag{31}$$

The action for the scalar field sector is taken to be,

$$\mathcal{A}_2 = \int d^4x[\mathcal{L}_\phi - \frac{1}{16\pi}\Sigma_{\mu\nu}\Sigma^{\mu\nu}], \tag{32}$$

where,

$$\mathcal{L}_\phi = \frac{1}{2}(\nabla_\mu\phi)^*(\nabla^\mu\phi) - \frac{\lambda}{4}(\phi^*\phi - \eta^2)^2, \tag{33}$$

and,

$$\Sigma_{\mu\nu} = \partial_\mu\chi_\nu - \partial_\nu\chi_\mu \quad , \tag{34}$$

It is well known that the ground state of this sector is described by the following solutions.

$$\phi_{vac}(x) = \eta e^{i\psi(x)} \quad , \tag{35}$$

and,

$$(\chi_\mu)_{vac} = 0 \quad . \tag{36}$$

It is evident from (28), (29) and (32) that $\mathcal{A}_0$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are invariant under local duality transformation respectively. Variation of the total action with respect to the particle trajectory y^μ_j and the complex 4-vector field a^μ leads to the following equations of motion :

$$\frac{dp^\mu_j}{d\tau_j} = \frac{\alpha_j}{2c}\,[\phi^*(y_j)G^{\mu\nu}(y_j) + \phi(y_j)G^{*\mu\nu}(y_j)]\,\frac{dy_{\nu j}}{d\tau_j}. \tag{37}$$

and,

$$\mathcal{D}_\mu G^{\mu\nu}(x) = \frac{4\pi}{c}\sum_j \alpha_j \phi(x)\frac{dy_j^\nu}{dt}\delta^3(\bar{x} - \bar{y}_j(t)), \tag{38}$$

respectively.

In the low energy limit, (37) and (38) are equivalent to the usual Lorentz force equation and ordinary Maxwell equations, respectively. This is because, in the vacuum configuration (35), the gauge covariant derivative $\mathcal{D}_\mu$ takes the form,

$$\mathcal{D}_\mu = \partial_\mu - i\partial_\mu\psi \quad . \tag{39}$$

By virtue of (19), under a local duality transformation the phase ψ transforms as,

$$\psi(x) \to \psi'(x) = \psi(x) + \theta(x) \quad . \tag{40}$$

Since the entire theory is invariant under local duality transformation, we are free to choose a gauge $\theta(x)$ = - $\psi(x)$ so that $\psi'(x) = 0$ because of (40). This immediately makes the gauge covariant derivative (in the new gauge) reduce to ordinary partial derivative (see (39)),

$$\mathcal{D}'_\mu = \partial_\mu \quad . \tag{40}$$

Furthermore, in this gauge the electromagnetic charge of the j-th particle is given by,

$$Q'_j = \alpha_j\phi'(x) = \alpha_j\eta \quad , \tag{41}$$

implying that the charges for all particles are constant and are real (corresponding to electric charge alone). It is easy to see making use of (40) and (41) that (37) and (38) reduce to,

$$\frac{dp_j^\mu}{d\tau_j} = \frac{1}{c}(q_e)_j F^{\mu\nu}\frac{dx_{\nu_j}}{d\tau_j}, \tag{42}$$

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} j_e^\nu, \tag{43}$$

and,

$$\partial_\mu \tilde{F}^{\mu\nu} = 0, \tag{44}$$

where,

$$(q_e)_j = Q'_j, \tag{45}$$

and,

$$j_e^\nu = \sum_j Q'_j \frac{dy_j^\nu}{dt} \delta^3(\bar{x} - \bar{y}_j(t)). \tag{46}$$

Equations (42) - (44) are the usual Maxwell-Lorentz equations in the absence of magnetic monopoles. Thus, in the low energy region the electromagnetic sector of this theory is identical to the conventional classical electrodynamics.

Before ending this section, we wish to draw attention to an additional local symmetry of the theory. Consider the following transformation,

$$a_\mu \to a'_\mu = a_\mu + \partial_\mu \beta(\phi^*), \tag{47}$$

where β is any complex differentiable function of ϕ^*. It can be easily shown that (47) leaves $G_{\mu\nu}$ invariant, and causes the action (29) pick up just boundary terms. Thus, equations of motion are left invariant under the transformation (47).

4. Summary and discussion

Most symmetries in nature are local symmetries e.g. gauge symmetries in electrodynamics and electro-weak theories, general covariance in Einstein's theory of gravitation, etc. It is therefore interesting to study the consequences of a local duality symmetry. Gauging this symmetry requires invoking a complex scalar field $\phi(x)$, that exhibits spontaneous breaking of duality symmetry. In the low energy limit, there is a gauge in which this theory automatically leads to conventional electrodynamics without magnetic charges. However, in the high energy domain, one expects new predictions that may be used to distinguish between the Maxwell electrodynamics and our model.

Acknowledgements

It is a pleasure to thank Professors R. Brandenberger, A. E. Everett, L. Ford, N. Mukunda, J. V. Narlikar, E. C. G. Sudarshan and A. Vilenkin for useful discussions. This work was supported by the C.S.I.R. fellowship.

References

1. P.A.M. Dirac, Proc. Roy. Soc. (London) A133 (1931) 60.
2. G. 't Hooft, Nucl. Phys. B79 (1974) 276; B105 (1976) 538.
3. A. Polyakov, JETP lett. 20 (1974) 194.
4. B. Felsager, Geometry, particles and fields (Odense University Press, 1987) p-475.

SECTION 3

GAUGE FIELDS IN CONDENSED MATTER

Quantum Phase in Action

A. ZEE

Institute for Theoretical Physics
University of California
Santa Barbara, California 93106-4030
and
Physique Theorique
Ecole Normale Superieure
75231, Paris

We are gathered here to pay homage to the quantum phase. Out of classical nothingness something quantum emerges.

One of the deepest mysteries in physics is the existence of two kinds of particles, bosons and fermions. We now know that in $2+1$ dimensional spacetime there are also anyons, such that when two anyons are exchanged, the wave function acquires a phase. In particular, when two semions are exchanged, the wave function changes by a factor of i. $2+1$ dimensional spacetime is not just less of a good thing compared to $3+1$ dimensional spacetime. It is homotopically different: a new physical concept, that of "going around", appears. It makes sense to say that a particle has gone around another. This basic fact is what makes the notion of anyons and fractional statistics possible.

Shortly after Wilczek, and earlier, Leinaas and Myrheim, proposed the existence of anyons, the question naturally arose as to whether these hypothetical particles can be incorporated into quantum field theory. The answer is yes, and the concept of gauge potential enters naturally. One simply couples a gauge potential to a conserved current of interest, and have the dynamics of the gauge potential governed by the Chern-Simons term.[1] In Maxwell dynamics, the spactime derivatives of the gauge field are related to the current. In Chern-Simons dynamics, the gauge field is directly related to the current. Life is simpler because one doesn't have to solve any partial differential equations. This is possible in $2+1$ spactime. In any dimensions, the current is of course a vector, the gauge field an antisymmetric tensor, but in $2+1$ dimensions, an antisymmetric tensor is also a vector, thanks to the Levi-Civita antisymmetric symbol.

This means that a charged particle would have a magnetic flux attached to it. Here the terms electric charge and magnetic flux refer of course to the quantities associated with the gauge potential we have introduced and not to the quantities studied by Coulomb, Faraday, Oersted and their friends. Long ago, Aharonov and Bohm told us that when a charged particle goes around a flux tube, the wave function acquires a phase. Thus if we have particles carrying both charge and flux, then when one such particle goes around another, the wave function acquires a phase. Fractional statistics is just a slice of the Aharonov-Bohm effect. Thus, two of the greatest names in physics meet two of the greatest names in mathematics.

In hindsight, this connection between Aharonov-Bohm and Chern-Simons appears so natural and so obvious that some workers in this field now think that it was known since the beginning of time. In fact, this connection only became clear in the fall and winter of 1983.

Over the last ten years, there have been many interesting applications using this formalism. Here I would like to talk about a recent discussion of tunnelling effect in double layered Hall systems.[2]

In this formalism, in the quantum Hall effect electrons are coupled to gauge potentials obeying Chern-Simons dynamics. As explained above, the electrons then carry magnetic flux. In a special state in the double-layered quantum Hall system (technically this corresponds to a certain matrix having a zero eigenvalue so that one of the gauge potential is liberated from being governed by Chern-Simons dynamics), the electrons in layer 2 act like flux tubes carrying flux -2π to the electrons in layer 1. Thus, an electron in layer 1 does not see the magnetic field imposed by the experimentalist, but an effective magnetic field equal to the magnetic field imposed by the experimentalsit minus 2π times the local density of electrons in layer 2. Now consider a long wavelength density wave such that as the density of electrons in layer 1 goes up the density of electrons in layer 2 goes down correspondingly. But then the effective magnetic field seen by the electrons in layer 1 also goes up. Thus, things can be arranged to work out very neatly. Even as the density of electrons in layer 1 goes up and down, those electrons can be made to believe that they are still just filling the first Landau level, not one too many, not one too few. Similarly, the electrons in layer 2 are also living under the illusion that they are filling just the first Landau level. Thus, as the wavelength of the density fluctuation goes to infinity, the energy cost of the fluctuation goes to zero. This is the physics behind the appearance of a gapless mode: the gaplessness is a consequence of an exquisitely balanced cooperation between the electrons in layer 1 and layer 2. The same physics is in fact responsible for anyon superfluidity. Technically, the gauge field liberated from being governed by Chern-Simons dynamics is now happily massless and governed by Maxwell dynamics.

The appearance of a gapless mode is consistent with symmetry considerations. In the absence of tunnelling, there are two separate $U(1)$ symmetries, corresponding to the conservation of the sum and difference of the electron numbers in the layers. In the special state described above, the $U(1)$ corresponding to the conservation of the difference of the electron numbers in the two layers is spontaneously broken and thus we expect a Nambu-Goldstone gapless mode.

Tunnelling, that is, interlayer hopping, corresponds to the explicit breaking of this $U(1)$ symmetry and thus according to general considerations, the Nambu-Goldstone boson becomes pseudo and acquires mass.

In the present formalism, the current describing the difference of the currents in the two layers is written as a curl of a gauge potential. When an electron tunnels from one layer to the other, this current is no longer conserved. When the divergence of the curl of a gauge potential does not vanish, we know that there is a magnetic monopole lurking in the vicinity. The spacetime integral of the magnetic flux coming out of the monopole is the spacetime integral of the divergence of the current, and hence the change in the difference of numbers of electrons in the two layers, equal to ±2 in the tunnelling event. Thus, the monopole in our formalism is quantized a la Dirac because electrons are discrete.

Dirac quantization of magnetic monopoles represents of course another manifestation of the Aharonov-Bohm effect. Dirac obtained magnetic quantization by requiring that the Aharono-Bohm phase acquired by a particle going around his string vanishes. Indeed, Coleman explains Dirac quantization by arguing in reverse. He describes a prankster trying to trick an experimentalist into believing that he or she has found the fabled magnetic monopole. The prankster introduces an arbitarily thin flux tube into the lab. The experimentalist can detect the flux tube by letting a charged particle move around and measure the resulting Aharonov-Bohm phase. It is precisely when the flux going through the tube is such that the monopole has the Dirac magnetic charge that the flux tube becomes undetectable. The experimentalist can then become very excited and proclaim the discovery of the magnetic monopole.

Thus, we have a Euclidean 3-space filled with a plasma of magnetic monopoles and anti-monopoles. Wherever there is a monopole, an electron tunnels from layer

1 to layer 2 at the corresponding point in spacetime. Whereever there is an antimonopole, an electron tunnels back from layer 2 to layer 1. Now we get to re-live the golden days of quantum field theories. One of the most celebrated results of the 1970's was the realization by Polyakov that in the presence of a dilute plasma of magnetic monopoles the photon acquires a mass.

This is completely consistent with our expectation from sysmmetry considerations. To summarize, we have the following "life story" of a gauge quantum. When it was governed by Chern-Simons dynamics, it was massive. After being liberated into a life of Maxwell dynamics, it becomes massless. But then non-perturbative tunnelling effects made it massive again, Technically, the plasma of monopoles is a Coulomb gas, and a Coulomb gas can be represented by a sine-Gordon theory. Expanding the cosine in the Lagrangian to quadratic order, one sees immediately that the sine-Gordon field is massive.

For his purposes Polyakov did not have to exploit the fact that the sine-Gordon field is in fact an angular order parameter. But we know that there is very interesting physics associated with angular order parameters! Incidentally, the order parameter is angular precisely because the magnetic monopole is quantized by Dirac. Wen and I are thus led to make the perhaps a priori rather surprising prediction that when a DC voltage V is applied across a double-layered Hall system, for certain special filling factors, there is an oscillating tunnelling current. In a word, there is a superfluid lurking in the system and hence there is Josephson-like current. Note however that the frequency is only half of the Josephson frequency because we don't have pairing here. We may entertain the hope that this effect will be experimentally detectable in the near future.

I hope to have conveyed the impression that the circle of theoretical ideas appearing in this subject are among the deepest in theoretical physics.

We encounter here quantum statistics, homotopic property of space, gauge potential, Chern-Simons and Maxwell dynamics, Aharonov-Bohm phase, Dirac quantization of magnetic monopole, quantum tunnelling, Nambu-Goldstone bosons, cooperative density and flux fluctuation and anyon superconductivity, discreteness of the electron, Coulomb gas, angular order parameter, and Josephson ocsillation. In the end, we can attribute all these strikingly beautiful notions to the fact that when we move from classical physics to quantum physics the complex number mysteriously appears on the scene.

With your indulgence, I will end by entertaining a speculation, in fact the same speculation[3] I made here in South Carolina a few years ago at another conference celebrating the Aharonov-Bohm effect. The appearance of statistics in quantum physics is one of the deepest mysteries in physics and in some ways is responsible for the current difficulties in particle theory. As Weisskopf discovered ages ago, fermions are nice and bosons are nasty. The self energy of a boson diverges quadratically. It is partly to cure this problem that supersymmetry was invented, to solve the so-called naturalness problem. We all know down what glorious paths supersymmetry has taken particle physics: from supersymmetry to supergravity to superstrings to supermathematics to superphysicists. Might it not be possible that quantum statistics is a composite notion? In the end, there are only fermions (or perhaps, only bosons.) After all, we know that a bound state of a boson and a magnetic monopole is a fermion (and vice versa.) There is an additional phase when two such bound states are interchanged. Indeed, it is possible to obtain reasonable quantum numbers for the observed quarks and leptons.[4] Some of the theoretical ideas I listed above are so deep that they ought to have further consequences for particle physics as well as for condensed matter physics.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation under Grant No. PHY89-04035. We also thank the theory group at the École Normale Supérieure for its warm hospitality.

REFERENCES

1. F. Wilczek and A. Zee, *Phys. Rev. Lett.* **51**, 2250 (1983).
2. X.G. Wen and A. Zee, *Phys. Rev. Lett.* **69**, 1811 (1992); *Phys. Rev.* **B47**, 2265 (1993).
3. A. Zee, in *Quantum Coherence*, edited by J. S. Anandan (World Scientific Publishing).
4. F. Wilczek and A. Zee, *Phys. Rev.* **D25**, 553 (1982), see section IV.

LIBERATING EXOTIC SLAVES

FRANK WILCZEK[†]
School of Natural Sciences, Institute for Advanced Study. Olden Lane
Princeton, New Jersey 08540, USA

ABSTRACT

The introduction of confined, "slave" fields is frequently useful as a formal device in models of condensed matter physics; it becomes a conceptual necessity for describing possible phases of matter where the slaves are liberated. Here I discuss some aspects of the fractional quantum Hall effect from this point of view, emphasizing analogies with phenomena in other areas of physics, particularly to the Meissner and Higgs mechanisms, and to confinement-deconfinement transitions. In this application, and in some recent attempts to model the normal state of copper oxide superconductors, it is important to employ slave anyon fields.

I have long admired Yakir Aharonov's style in physics: to continue to puzzle over that which is intrinsically strange, even in domains where more jaded spirits have lost, from mere familiarity, their sense of wonder. This child-like quality has led him to make fundamental discoveries where few would anticipate that fundamental discoveries could still be made, and--as we all must acknowledge on this occasion--it obviously has kept him young!

In that spirit, I hope, I would like to discuss with you today a personal perspective on the fascinating complex of new states of matter forming the "quantum Hall complex," which I have developed in response to some simple puzzles that have bothered me for a long time. One of the puzzles, as I shall describe momentarily, has to do with gauge invariance. The other is broader: is the fractional quantized Hall effect as special and isolated as it seems at first sight, or can its occurrence be related to other deep ideas in theoretical physics? I have found my perspective quite comforting and informative, and I think it is different at least in emphasis and some significant details from what has appeared in the literature (including my own work.) However, I must quickly add that it in no way alters with Laughlin's basic physical picture of an incompressible quantum liquid, nor will it be used here to derive new results that could not be found otherwise.[1-3]

1. Critique of Laughlin's Quantization Argument

1.1. The Argument

Shortly after the experimental discovery of the integer quantized Hall effect,

[†] Research supported in part by DOE grant DE-FG02-90ER40542

Laughlin[4] proposed an argument, based on gauge invariance, that explains why the conductance is quantized. The argument proceeds from the physical hypothesis that in the conditions where the quantized Hall effect is observed the electrons form an incompressible fluid in the bulk, to show that the conductivity of the fluid (to be defined, in a precise geometry, momentarily) must be an integer multiple of a certain combination of fundamental constants, *viz.* e^2/h. With some important refinements due to Halperin,[5] this argument remains the foundation of the theory of the effect. I would like briefly to recall its essence.

Imagine an annulus containing electrons held at low temperature and subject to a large perpendicular magnetic fields, and such that the inner and outer edges are connected by an ordinary wire and held at a voltage difference V. Suppose that we have the conditions of the quantized Hall effect, that is, by hypothesis, that within the bulk of the annulus there is a incompressible electron fluid. This means that there is, for each value of the current circulating around the annulus, a unique bulk state of minimum energy. It can be constructed, locally, from the unique, isolated ground state by a Galilean transformation.

Now let us suppose that there is a current I circulating around the annulus, and consider the effect of switching on one quantum h/e of flux in the void within the annulus. At the end of this operation we have produced a gauge field, that (for electrons within the annulus) is gauge equivalent to zero. Thus the bulk state, assumed unique, must return to its original form. The only change that can have occurred, is that some electrons from one edge might have been transferred to the other edge, through the wire.

We can calculate the work done during this operation in two different ways. On the one hand, we have transferred some charge ne through a voltage V; thus the work is neV. On the other hand while the flux is being increased there is an azimuthal electric field, which does work on the circulating current. One easily computes in this way that the work done is $(h/c)I$. Upon equating these, one finds for the conductance:

$$V/I = ne^2/h\,. \tag{1}$$

Thus, this transverse conductance is quantized in terms of fundamental physical constants.

A slight variant of this argument corresponds less well to a practical experimental set-up, but is perhaps simpler conceptually and will be useful for my later purposes. Consider the same geometry and the same process of cranking on flux, but now with no transverse current and no voltage. As the flux is turned on, again some integer k number of electrons is transported. There was an azimuthal electric field as the flux was turned on, and thus, for a determinate transverse conductivity, a radial current. The electric field is proportional to the time rate of change of the flux, so over the course of turning on one quantum of flux there is a definite integrated radial current, or in other words a definite charge transfer. Equating this charge transfer to ke, one finds the same quantization condition on the transverse conductivity as before.

1.2. Too Good to be True?

The Laughlin quantization argument is so simple and beautiful, and so directly addresses the central phenomenon, that one cannot seriously doubt its essential correctness. Unfortunately, it is *too* good. Shortly after it was proposed and digested, experimentalists discovered states where the conductance is quantized, but now as a definite fraction of e^2/h rather than as an integer multiple. These states occur when the density is close to (the same) definite fraction of the density corresponding to a full Landau level. The jargon here is that there is a plateau in the resistivity around filling fraction $\nu = \rho/(eB/\hbar c)$; meaning that when the ratio of density to magnetic field is close to this value the conductivity remains at the quantized value $\nu e^2/h$. The first discovered and most robust such state (as reflected in the width of the associated plateau and the allowed range of impurities and temperatures) occurs at $\nu = 1/3$. For simplicity and concreteness I shall mainly focus the discussion on that state, although by now quantized Hall states at many other fractions have been observed and there is a beautiful, extensive theory of them—in fact several such theories.[6]

Now we seem to be in the embarrassing position, with the preceding gauge invariance arguments, of having proved too much. The conductance is not quantized in integers times e^2/h for incompressible bulk states, after all. What has happened?

1.3. The Microscopic Perspective

There is a successful microscopic theory of the fractional quantized Hall effect. So before I get carried away with grandiose rhetoric about breaking and amending gauge invariance, it behooves me to demonstrate how one understands at a "mechanical" level how the general gauge invariance argument, which seems so clear-cut in leading to integer quantized conductance, develops the necessary subtleties in the microscopic theory.

1.4. Lightning Review of Incompressible Hall States

As we have already seen in our discussion of the integer effect, the quantized conductance is a fairly direct manifestation of the existence of an incompressible quantum fluid. That is, the electron fluid has a preferred density pinned to the value of the external magnetic field. There must be an energy gap to deviations from this preferred density: such deviations must be accommodated by localized inhomogeneities, rather than in arbitrarily long wavelength "sound waves" which—if they existed—could have arbitrarily small energy. In the case of the integer quantized Hall effect the preferred density simply corresponds to filling an integer number of Landau levels, and the gap is quite easy to understand. Indeed, to raise the density *here* and lower it *there* we must excite a particle to the next Landau level *here*, which costs a finite minimum amount of energy equal to the splitting between Landau levels, that is not compensated by allowing a hole *there*[†].

[†] The lowest energy density fluctuations actually occur at a *finite* wavevector. These excitations,

Laughlin himself[8] was quick not only to recognize the physical meaning of the new observations, but also to propose a rationale for why specific special (non-integer) filling fractions should be preferred. Let me very briefly recall the main points, since I shall want to build on them.

First I need to remind you of some basic results about electrons in a strong magnetic field (here, as throughout, I am assuming that the motion of the electrons is confined to a plane.) The energy levels are highly degenerate Landau levels, with a density of states $2\pi/l^2$ per unit area per Landau level, where the magnetic length l is defined through $l^2 \equiv eB/\hbar c$. The splitting between levels is $\hbar$ times the cyclotron frequency, *viz.* $\Delta E = \hbar(eB/mc)$. At low temperatures and for densities small compared $2\pi l^2$ it ought to be a good approximation to restrict attention to states formed from single-particle states confined taken from the lowest Landau level, unless there is some very special energetic advantage to admixing higher levels (so as to minimize the interaction energy.) Within the lowest Landau level, the single particle wave functions take a particularly attractive form if one employs the so-called symmetric gauge, defined by the vector potentials $A_x = By/2$, $A_y = -Bx/2$. With this gauge choice, the wave functions in the lowest Landau level take the form

$$\psi = f(z)e^{-\frac{1}{4}|z|^2} \tag{2}$$

where $f(z)$ is an arbitrary analytic function of $z \equiv x + iy$, subject to a reasonable growth condition so that the wave function is normalizable, and distances are measured in units of the magnetic length. A basis of orthogonal vectors in this Hilbert space is provided by the functions with $f_l(z) = z^l$. l is the canonical angular momentum around the origin, which here is intrinsically non-negative. For reasonably large l, the corresponding wave function is concentrated in a circular ring of radius $\sqrt{2l}$ and width $\sqrt{2\pi}$ around the origin. It follows, by comparing the size of the region where the wavefunction is large to the inverse density, or by direct calculation, that the supports of these wave functions are highly overlapping.

Now let us consider an assembly of (non-interacting) electrons. Let us suppose that they subject to a very small potential that draws them toward the origin, but does not appreciably change the form of the wave functions (that is a second order effect). Then the ground state will be composed out of the wave functions with the smallest values of l, consistent with Fermi statistics. It will be the Slater determinant

$$\psi_1 = \det\{z_r^{c-1}\}e^{-\frac{1}{4}\sum|z_k|^2} , \tag{3}$$

where the row variable r, the column variable c, and k all run from 1 to N, the number of electrons. Given the spatial character of the wavefunctions as discussed

the so-called magnetorotons[7] can be regarded, intuitively, as bound states of quasiparticles and quasiholes. They therefore bear a family resemblance excitons in semiconductors; however unlike most excitons they do not easily cascade down and annihilate, because semiclassically the Coulomb attraction between them—in the presence of the strong ambient magnetic field—causes a drift in the perpendicular direction, and thus induces orbital motion. Of course the magnetorotons, unlike the quasiholes and quasiparticles discussed below, carry no net charge.

above, one easily realizes that ψ_1, for large values of N, represents a droplet of uniform density 2π and radius $\sqrt{2N}$, with some fuzziness in an annulus of width unity near the edge. For later reference let me also record the Vandermonde identity

$$\det\{z_r^{c-1}\} \;=\; \prod_{k<l:k,l=1}^{N} (z_k - z_l) \tag{4}$$

Now Laughlin's inspiration was to notice that the cube of this wave function has remarkable qualities, that make it a particularly attractive trial wave function for an assembly of interacting electrons. The Gaussian factor is then not appropriate for the lowest Landau level, but this can be compensated by a trivial redefinition of the length unit, which we suppose done. Then clearly one has a wavefunction again describing a uniform droplet centered at the origin, now with radius $\sqrt{2N/3}$, density $2\pi/3$ (that is, filling factor $1/3$) and fuzziness in an annulus of width $1/\sqrt{3}$ after the rescaling. The Laughlin wave function is particularly advantageous if the electrons have repulsive short-range interactions, because it enforces a triple zero as one electron approaches another. A large number of numerical studies have shown that it is a very good representation of the ground state wave function, for a variety of repulsive interactions.

From a physical point of view, the most remarkable thing about the Laughlin wave function (and its various generalizations – see below) is its rigidity. It picks out a particular filling factor in the bulk. Deviations from this average density will have to be accommodated by localized disturbances. As we shall make much more precise below, the situation is analogous to what one has for type II superconductors, where magnetic fields are not allowed in the bulk, but can penetrate only in localized vortices. Laughlin proposed a form for these disturbances, that compares very well with numerical and experimental data. It is that a minimal quasihole localized around z_0 is represented by multiplying the wave function with a factor that pushes electrons away from z_0 by adding one unit of angular momentum around that point.

$$\text{quasihole factor} \;=\; \prod_1^N (z_k - z_0)\ . \tag{5}$$

This gives a density deficit; there is an analogous but slightly more complicated construction for an enhancement, the quasiparticle. There is an important *gedanken* production process for the quasihole: it is what you get by adiabatically switching on one unit of magnetic flux at z_0. The quasiholes are rather exotic: they carry fractional charge and fractional statistics. These properties can be shown directly from the microscopic theory.[9] I will forego that pleasure here, however the result will be central to our later considerations.

1.5. The Gauge Argument, Reconsidered

With this background, let us return to the gauge invariance argument. The second form of the argument is a little easier to discuss, so let's consider it.

There appears to be a technical awkwardness at the outset, in that we would like to work in an annular geometry for the fluid and to include some mechanism for taking electrons in one side and out the other, whereas the simple wave functions are for a droplet geometry. Fortunately there is a way around this that is quite simple and instructive for our purposes. We have already mentioned that wave functions with a high power z^l times the usual exponential $e^{-\frac{1}{4}|z|^2}$ are concentrated in a small ring of radius $\sqrt{2l}$ and width $\sqrt{2\pi}$ around the origin Thus to put a hole in the droplet of radius R, and produce an annulus of quantized Hall fluid, we should multiply the wave function by a factor

$$\text{Annulizing factor} \; = \; \prod_k z_k^{(R^2/2)} \; . \tag{6}$$

Now you will not fail to notice that the annulizing factor is nothing but $R^2/2$ quasiholes at the origin. A large number of quasiholes do literally make a (classical, spatial) hole in the fluid! Also, since the quasiholes are the end result of adiabatic insertion of a unit of magnetic flux—that's how we (following, of course, Laughlin) constructed them—we conclude that adiabatic insertion of flux drills a hole in the droplet.

Although it is somewhat off the point for this talk, it is quite interesting and appropriate to the occasion to note that *by redistributing flux that lies entirely in the empty void within the fluid annulus, one changes the shape of the annulus.* Thus some of the factors of $\prod z$ in the annulizing factor could be changed to $\prod(z - \alpha)$. This is a truly remarkable example of an Aharonov-Bohm type effect, in my opinion. That is, although one has "pure gauge" outside the flux tube, by moving the tube around one produces definite physical effects. (There is a pedestrian explanation for this—the moving flux tube produces an electric field at distant points.) The dynamics of motion within this manifold of quasi-degenerate states, produced by moving flux in the void, is governed by the theory of edge excitations. Perhaps it is even a practical proposition to produce these excitations by manipulating flux in this way. (End of digression.)

So now we should be able to see, in the microscopic theory, how it can be that the gauge invariance argument becomes subtle, in such a way that inserting a single unit h/e of flux does not transport an integral number of electrons—while inserting three units does.

It is really quite simple and beautiful. The point is that when the power in the annulizing factor is a multiple of three, we can again write the wavefunction in Vandermonde-Laughlin form. That is (stripping away the Gaussian factors):

$$\begin{aligned} \prod_{k=1}^{N} z_k^{3L} \prod_{(k<l):k,l=1}^{N} (z_k - z_l)^3 \quad &= \\ \prod_{k=1}^{N} z_k^{3L} (\det\{z_r^{c-1}\})^3 \quad &= \\ (\det\{z_r^{c+L-1}\})^3 \; , & \end{aligned} \tag{7}$$

where one has $N \times N$ determinants with row index r and column index c. Thus to change L by one unit, to $L+1$, we need only to change the wavefunction of one electron, changing a z^L to a z^{L+N}. In physical terms, this means removing an electron from the inner edge and transporting it to the outer edge. (Note that the *minimum* occupied level has been emptied, and the *minimum* available unoccupied level has been filled.) That is the sort of operation an ordinary wire is happy to do. The remaining electrons in the annular drop can be entirely passive, and need not re-arrange their correlated wavefunctions.

It is quite a different story if you change the flux by one unit. That does not correspond to transport of an electron from the inner edge to the outer edge, leaving the bulk intact. Indeed, as we have just seen, the latter operation in its minimal form unambiguously corresponds to changing the flux by *three* units. The physical operation that corresponds to one flux unit, is creation of a quasihole-quasiparticle pair at the inner edge, followed by transport of the quasiparticle to the outer edge. This is not an operation an ordinary wire will do for you. There is an amplitude for it to occur by the quasiparticle tunneling across the sample, but since it requires a simultaneous rearrangement of all the electrons this amplitude will be exponentially small. In the thermodynamic limit of an infinite number of electrons, at zero temperature, it will not occur at all. Then we are justified in saying that gauge invariance has been spontaneously violated, in the only sense it ever is: while the gauge transformation with three flux units connects one *accessible* state to another, and represents a legitimate symmetry; but the transformation with a single flux unit, although formally valid, is useless because it relates amplitudes for processes in our world only to amplitudes for processes in another, inaccessible one.

2. Introducing, and Liberating, Confined Slaves

2.1. Analogies of iQHE[§] *and Superconductivity*

One cannot long reflect on the properties of the incompressible Hall states without noticing many analogies between their properties and those of ordinary superconductors. Let me mention a few of the most striking ones:

- In the quantum Hall system, there is a vanishing longitudinal resistivity. Thus the current flow is non-dissipative, as in a superconductor. Strictly speaking, this is true only at zero temperature. However, this fact does not spoil the analogy: we are dealing with a two-dimensional system, and in two dimensions the superconducting transition is also at zero temperature. Indeed, the reason is the same in both cases: there is a finite energy gap to vortex production, which leads to finite though exponentially small dissipation at any non-zero temperature.
- In both cases, one has an energy gap to charged excitations.
- In both examples, one has *rigidity against an applied magnetic field.* In the case of superconductors this is of course the famous Meissner effect, but it may seem

[§] I shall use this notation for the incompressible quantum Hall effect, which is a mouthful. The lower case i is used here, because IQHE is already used to indicate the integer quantized Hall effect.

to be a rather peculiar thing to say about iQHE states, since they occur immersed in a magnetic field from the start. Nevertheless they exhibit a form of rigidity, in that changes of the field away from a preferred value, pinned to the effective density, are disfavored. Here by effective density I mean the nominal density as given by the Hall coefficient, which is constant over a given plateau – in the analogy, we could call this the superfluid density.

- In both cases, one has vortex-like objects. We have of course just seen this in our discussion of the iQHE, where the quasiparticles are in some sense vortices, and it is a famous fact for type II superconductors.

- In this vein, there is also the analogy that the non-dissipative state requires that the vortices be pinned. The pinning is much easier in the iQHE case, because the vortices are electrically charged and subject to a large magnetic field, so they will be happy to make closed orbits on electric field equipotentials. (Nevertheless some impurities must be present to make these equipotentials form closed lines, or else there will be no plateau. Indeed for a translationally invariant system the Hall constant must be equal to the carrier density, by Galilean invariance, and it cannot "stick" at a preferred value as the density varies.) At finite density the quasiparticles would presumably, given their large effective band mass and repulsive interactions, form a Wigner crystal, analogous to the Abrikosov flux lattice.

On the other hand one has the apparent contrast, that the iQHE states but not ordinary superconductors support exotic charge and statistics for the quasiparticles. Also, as I discussed in the first part of this talk, the breaking of gauge invariance is rather different in the two cases. For an ordinary superconductor, the periodicity in the Aharonov-Bohm type *gedanken* experiments we considered there would be $h/2e$ instead of the $h/(e/3)$ we encountered for the $\nu = 1/3$ state. The difference is profound: whereas in the first case one has a higher degree of flux-periodicity (that is, a smaller flux quantum) than might of been anticipated, reflecting a pairing order parameter, in the later case one has a subharmonic periodicity.

2.2. Introducing Exotic Slaves

The subharmonic periodicity in flux coexists, in the iQHE, with the existence of fractional charge, and one would like to think that there is an organic connection between them. Such a connection will arise, similarly to what one has in superconductivity, if one requires that the integral

$$\begin{aligned} \text{charge transport phase} &= e^{iq\oint A_\phi d\phi} \\ &= e^{iq\Phi}\,, \end{aligned} \tag{8}$$

describing the phase acquired by a particle of charge q transported around a closed loop enclosing flux Φ to be unity, for a *fractional* charge $q = e/3$. This single-valuedness, in turn, will have to be imposed if there is condensation of a field with charge $e/3$. The case for an organic connection thus becomes compelling. For the existence of fractionally charged quasiparticles supplies, on the face of it, a natural candidate for the desired condensate field: namely, of course, the field ψ that creates the fractionally charged quasiparticles.

There is a difficulty, however. If ψ is to condense one would like it to be bosonic. But that desire appears to conflict with another: one would also like to be able to have possibility for an electron decay into three identical quasiparticles. For the quasiparticles are supposed to be the important charged low-energy excitations, and this is the minimal decay channel that allows an electron to communicate with them, while conserving charge. Clearly, if the quasiparticles are bosons this decay is not going to be possible. One needs particles with exotic *anyon* quantum statistics, in order that a state of three identical particles can have the quantum numbers of a fermion. Furthermore the microscopic theory teaches us that the quasiparticles are in fact anyons, and an electron can in fact decay into three of them. (Another possibility would have been to have more than one kind of quasiparticle: for example, one could reproduce the electron quantum numbers if there were in addition a light neutral fermion excitation, so that an electron could decay into three identical bosons and the neutral fermion. There may be iQHE states with this kind of non-minimal structure—a candidate $\nu = 1/2$ state of this kind has been described.[10] However for the more conventional iQHE states, a minimalist procedure works out quite elegantly, as we shall see.)

So we seem to have arrived at a dilemma: on the one hand we want to have a bosonic field to create the quasiparticles, so that the field can condense; but on the other hand we want the quasiparticles to be anyons, so that they can reproduce the electron's fermion statistics. Fortunately, these requirements only appear to be contradictory. Theoretical work on quantum statistics in 2+1 dimensions has shown that a bosonic field, properly coupled to a gauge field, can create anyons of any type.[3] The way of this is done is called the Chern-Simons construction. It works as follows. One couples the field ψ using the minimal coupling procedure to a gauge field a that does not have an ordinary Maxwell kinetic energy term, but instead only a "Chern-Simons" term

$$\Delta\mathcal{L}_{\rm CS} = \frac{n}{4\pi}\int \epsilon^{\alpha\beta\gamma} a_\alpha f_{\beta\gamma} \,. \tag{9}$$

Now one can demonstrate, without much difficulty, that the quanta produced by ψ will have their quantum statistics altered, by the presence of the so-called Chern-Simons gauge field a of which they are a source. And—at least in the point particle limit, for which the concepts are clearly defined—this change in the statistics of the quanta is the only effect of coupling in as. This construction is therefore *a* valid, and *the* minimal, way of implementing statistical transmutation—that is, the creation of quanta of one statistics by fields with another.

I originally called fields such as a "fictitious" gauge fields. The newer terminology is in many ways preferable, but the old terminology did have the advantage of emphasizing that the a do not introduce new local degrees of freedom. One can in principle fix a gauge and solve for the as in terms of ψ. (The price for this is that the resulting action is complicated and no longer manifestly local.)

Although I do not intend to pause for a full demonstration here, it is especially appropriate on this occasion to note that the Aharonov-Bohm effect lies close to the heart of statistical transmutation. For the essence of the matter is that one finds, on solving the equations of motion for the gauge fields a, that the effect of the Chern-Simons coupling is simply to turn each quantum created by ψ into a source of flux, as well as charge. Indeed, on varying with respect to a_0 one finds the equation

$$\rho = -\frac{n}{2\pi} f_{12} \tag{10}$$

relating the particle number density to the Chern-Simons magnetic field. Note that in two space dimensions one has flux *points*, as opposed to the familiar flux lines, and one can properly speak of flux associated to a point particle. When one such particle circles around another the wave function acquires, as Aharonov and Bohm taught us, a phase proportional to the product of charge and flux. But such a phase is operationally indistinguishable from the effect of quantum statistics! And that's why one can freely change the statistics of the quanta created by a given field ψ by coupling ψ to a Chern-Simons gauge field.

We can summarize these considerations succinctly as follows. As far as the quantum numbers of charge and statistics are concerned, we can represent a field capable of creating an electron as

$$e \sim \psi\psi\psi\ , \tag{11}$$

where ψ is a *bosonic* field with electric charge $e/3$, properly coupled as well to a Chern-Simons gauge field. With our conventions, the correct choice is simply $n = 3$ in Eq. (9).

It has frequently been useful in condensed matter problems to introduce, as a mathematical device, representations of electron fields as products of other "slave" fields. One might, for example, represent the electron as a product of a neutral fermion "spinon" field and a charged boson "holon" field. As long as there is a constraint in place, forbidding the separate propagation of quanta of these fields, this is just a mathematical device. One is then in a confined phase, analogous to the confined phase for quarks in QCD. What we have done here is introduce a particular exotic kind of slave field, with fractional charge and statistics. As long as its quanta are kept confined--as might be implemented by a Z_3 gauge field coupling doing this is just a mathematical device. As long as we consider only scales much larger than the confinement scale, we will not have changed the physical content of the theory. The procedure will be useful if added flexibility introduced by the slave variables allows us to represent excitations or correlations that are awkward to describe (i.e. non-local) in terms of the original variables.

2.3. iQHE as a Modified Meissner Effect: Liberating the Slaves

We introduced the slave field ψ with two purposes in mind: the straightforward one, that after all there are quasiparticle states with exotic quantum numbers in the iQHE, so we should have fields to create them; and the deeper one, that we would like to have a condensation, or vacuum expectation value, of charge $e/3$ fields, so as to understand the subharmonic flux periodicity in the Laughlin argument.

Can ψ condense? At first hearing the idea might sound mad. After all ψ is a charged field, and the essence of the Meissner effect is that charged fields cannot condense in the presence of a background magnetic field. They are, in the jargon, frustrated. Since the iQHE necessarily takes place in a large background magnetic field, the proposed condensation sounds to be grossly anti-Meissner.

On deeper consideration, however, one discovers within this seeming difficulty the central point of this circle of ideas. Let us recall how one understands the Meissner effect, in the language of condensation. In the free energy associated with a charged condensing field η one has a gradient term

$$|\nabla\eta|^2 \;=\; |\partial_\mu\eta - iqA_\mu\eta|^2 \tag{12}$$

involving the gauge covariant derivative. Now a constant magnetic field introduces a vector potential A which grows with the distance, and whose effect, since it is solenoidal, cannot be cancelled by the ordinary derivative term, which is longitudinal. Thus to maintain a non-zero expectation value for the magnitude of η costs a free energy density which grows with the distance, and this can never be favorable.

Now in the analogous considerations for our exotic slave field η, we must include not only the electromagnetic gauge field but also the Chern-Simons field a. And then we realize, that there is a possibility for A and a to cancel, thus allowing for the possibility of a uniform condensate. This will occur when the part of $\frac{e}{3}A + a$ that grows with the distance cancels. That, in turn, requires that the average flux density associated with this combination of fields vanishes. In view of Eq. (10), this occurs when one has the relation

$$\frac{e}{3}B \;=\; b \;=\; \frac{2\pi}{n}\rho \;=\; \pi\rho_e\ , \tag{13}$$

where in the third equality we have taken into account the $n = 3$ demanded by quantum statistics, and that the quasiparticle density is three times the electron density. Thus the cancellation takes place precisely at filling fraction $\nu = 1/3$. Whereas the ordinary Meissner effect for a superconductor tends to exclude magnetic field, the modified Meissner effect taking into account the statistical transmutation, excludes *deviations* of the magnetic field from a fixed multiple of the density (and, of course, *vice versa*). Deviations from zero field in the superconductor, or from the desirable density in the iQHE, are accommodated most cheaply by allowing inhomogeneities—vortices in the first case, quasiparticles in the second. In fact the quasiparticles are vortices too—but in the Chern-Simons field, not the electromagnetic field. Only by allowing such inhomogeneities can one preserve condensation in bulk, which requires the integrated form of Eq. (13). That is the essence of the modified Meissner effect.

Another feature of the situation is that the condensation of ψ into a Higgs phase entails, as a consistency requirement, deconfinement of its quanta. One cannot, after all, confine vacuum quantum numbers! Thus the two purposes which motivated us to introduce the confined slaves, namely on the one hand to have fields which described the exotic quasiparticles once they are liberated, and on the other hand to have fields capable of condensation, are intimately related in their realization.

2.4. Past and Future

Well that concludes the main story I wanted to tell you today, and I think it is a very nice story as far as it goes. I hope I have conveyed how the concepts of fractional charge and statistics, the Chern-Simons construction of the latter, and the modified Meissner effect ineluctably come together in a coherent account encompassing both the iQHE and ordinary superconductivity. It does justice, I believe, to the 'paradoxical' nature of gauge symmetry in the fractional quantum Hall states that one encounters upon taking the Laughlin quantization argument seriously, as we discussed above.

This story has both a history and, I hope, a future. I'd like briefly to comment very briefly on these, although you should be warned that in neither case do I speak with authority.

Girvin[11] stressed the analogies between superconductivity and the iQHE very early on, made pioneering attempts to construct a consistent, unfrustrated order parameter, and recognized the importance of the statistical gauge field in this regard. Girvin and MacDonald[12] made an important connection to the microscopic theory. The early ideas were refined and extended in important ways by Zhang, Kivelson, and Hansson,[13] and by Read.[14] There is an interesting discussion of this body of work in Stone's book.[2]

In previous work, as far as I know, integrally charged condensates have been emphasized. For example in the approach of[13] one couples the statistical gauge field to the electron field to make it a "super-fermion"—though created by a bosonic field¶. This can be done with a Chern-Simons coupling $n = \frac{1}{3}$. With this value the modified Meissner argument gives the same relation between real magnetic field and electron density as was discussed above.

In this talk I have discussed how one is naturally led to the fractional charge condensate. Of course the existence of such a condensate does not contradict the existence of an electron condensate, but postulates additional structure. I think there are significant advantages to this point of view. For example the quantization of n in integers is required, for consistency, when one considers carefully the quantization of the Chern-Simons theory on topologically non-trivial surfaces. The appearance of integers multiplying the Chern-Simons term, and more generally (for

¶The notion of "super-fermions," that is of particles for which the wave function not only changes sign—that is, accumulates phase π –but accumulates phase 3π, say, may appear incoherent at first sight. After all, there is no denying that $e^{i\pi} = e^{3i\pi}$. However, it does have a concrete meaning *operating among states within the lowest Landau level.* For in that context the relative angular momentum must be positive, and the effect of boosting the angular momentum by two units is to change the spectrum of allowed values, so that the angular momentum has to be at least three. Without the positivity restriction on angular momenta that operates in the lowest Landau level the allowed spectrum would not be altered, and the notion of "super-fermion" would be quite dubious.

iQHE states at higher levels in the hierarchy) matrices of integers describing several coupled Chern-Simons theories, plays a crucial role in Wen's theory of edge states.[15] Thus both for understanding the accuracy of the quantization in the FQHE in a fundamental way, and for connecting ideas about the bulk state to the successful theory of edge states, it is important to have integers.

Having identified something like an order parameter, one might like to continue the analogy with superconductivity by considering inhomogeneous situations, response to external fields, and so forth, by solving classical equations using an effective Lagrangian, in the style of Landau and Ginzburg. In attempting this, however, one must recognize that the fields involved in such an effective Lagrangian cannot be regarded as normal local 2+1 dimensional fields, because they should only create and destroy quanta in the lowest Landau level (which makes them effectively 1+1 dimensional).

As a concrete example, one would like to use an effective Lagrangian to describe the motion of quasiparticles in response to slowly varying external electric and magnetic fields, or their scattering at small momenta. Indeed these most basic processes involving quasiparticles are perhaps the most fundamental observable processes governed by their exotic charge and statistics, so one would like to have an explicit description of them. Even in the simplest case of the integer quantized Hall effect, where the quasiparticles are the electrons themselves, it would seem that a more direct approach to calculating charged particle drifts in the lowest Landau level is appropriate, and this has quite a different flavor from solving simple classical field equations. This subject needs more work.[16]

2.5. Coda: Question of Statistics in Spin-Charge Separation

There are several indications that the normal state of the CuO high temperature superconductors, for the dopings at which they exhibit superconductivity, is an anomalous metal. Perhaps the most striking anomaly is the linear dependence of resistivity on temperature, down to quite low temperatures. This is different from what is expected for a Fermi liquid, even after allowing for various possible complications.[17,18] On the other hand there definitely are indications that a Fermi surface exists, at least in the sense that there is a significant singularity in the density of states (imaginary part of the electron Green function) at a surface in momentum space. However, the size of the Fermi surface appears in some classes of experiments, particularly photoemission, to be roughly normal; whereas Hall effect measurements, if interpreted as reflecting Fermi surface parameters, give a very different picture. Although these experiments are not entirely straightforward to interpret (because the Fermi liquid theory fails to describe their temperature dependence correctly, the foundations of the analysis are insecure), on the face of it they seem to indicate a small Fermi surface for small doping, with positive (hole-like) carriers. Thus they seem to reflect not the entire electron density, but rather its deviation from half filling.

Motivated by these and other experimental results, which appear to require a 2-component model, and by experience with 1+1 dimensional models, Anderson and others have proposed that the anomalous state is characterized by *spin-charge separation*, that is the existence of separate spin and charge degrees of freedom—spinons and holons. Electrons are supposed to decompose into these more basic objects. This is known to happen in 1+1 dimensions, even for very weak coupling.[19] In 2+1 dimensions the situation is much less clear. The infrared singularities that drive 1+1 dimensional metals, even for small coupling, to qualitatively different behaviors are substantially weaker in 2+1 dimensions.

Nevertheless one is motivated by the phenomenology, by the 1+1 dimensional models, and by the "existence proof" provided by the foregoing analysis of the iQHE, to consider the possibility that in the CuO materials the transition to the normal state involves a liberation of exotic slaves. If there are states of matter in 2+1 dimensions wherein electrons do separate into spinons and holons, the question arises what is the statistics of these particles. The most obvious assignment is boson statistics for one, fermion statistics for the other.[20] On closer examination however this assignment appears to lead to severe difficulties.[20] The Bose condensation temperature tends to be very high, and if it occurred it would lead to striking effects, none of which are observed. My colleagues and I suggest instead[21] to consider the possibility that both species are half-fermions. This avoids the Bose condensation problem. Recent work on gauge theories[22] inspired by the Halperin-Lee-Read[23] theory of the compressible Hall states near $\nu = 1/2$ suggests another advantage of assigning fractional statistics to the spinons and holons, namely that they lead to a pattern of anomalous behaviors at least qualitatively suggestive of CuO phenomenology. There is a nominal Fermi surface, but as one approaches the Fermi momentum there is a severe renormalization of the effective mass, so that the singularities and temperature dependences are not of the form predicted by Fermi liquid theory.

A detailed account of this work will be appearing shortly. I wanted to mention it here as it is so closely allied to the ideas discussed in the body of the talk, and perhaps gains some credibility from the association.

References

1. In keeping with the informal nature of this talk, only a few references are given. For general background material on the subject treated here I suggest R. Prange and S. Girvin, *The Quantum Hall Effect* (Springer-Verlag, New York, 1987) and the following two items.
2. M. Stone, *The Quantum Hall Effect* (World Scientific, Singapore, 1993).
3. F. Wilczek, *Fractional Statistics and Anyon Superconductivity* (World Scientific, Singapore, 1990).
4. R. Laughlin, *Phys. Rev.* **B23** (1981) 5632.
5. B. Halperin, *Phys. Rev.* **B25** (1982) 2185.
6. They may be termed the variational-numerical theory, the hierarchy theory, the Chern-Simons theory with Landau-Ginzburg and duality branches, the

composite fermion theory, and the adiabatic statistical transmutation theory. This list may be incomplete. I attempt to survey these various approaches, emphasizing the last, with adequate references in my talk at the 150 anniversary of Boltzmann's birth (Rome, 1994). It will appear in the proceedings under the title *Statistical Transmutation and Phases of Two-Dimensional Quantum Matter.*

7. S. Girvin, A. MacDonald, and P. Platzman, *Phys. Rev. Lett.* **54** (1983) 581.
8. R. Laughlin, *Phys. Rev. Lett.* **50** (1983) 1395.
9. D. Arovas, R. Schrieffer, F. Wilczek, and A. Zee, *Nuclear Physics* **B251** (1985) 117.
10. B. Halperin, *Helv. Physica Acta* **56** (1983) 75.
11. S. Girvin, in reference 1.
12. S. Girvin and A. MacDonald, *Phys. Rev. Lett.* **58** (1987) 1252.
13. S. Zhang, H. Hansson, and S. Kivelson, *Phys. Rev. Lett.* **62** (1989) 82.
14. N. Read, *Phys. Rev. Lett.* **62** (1989) 86.
15. Reviewed in X.-G. Wen, *Int. J. Mod. Phys.* **B6** (1992) 1711.
16. R. Levien, C. Nayak, and F. Wilczek, paper in preparation.
17. E. Abrahams, *Beyond the Fermi Liquid*, Rutgers preprint (unpublished, 1993).
18. C.Varma *Theory of the Copper-Oxide Metals*, Bell reprint (unpublished, 1994).
19. F. D. M. Haldane, *J. Phys.* **C14** (1981) 2585.
20. P. Lee and N. Nagaosa, *Phys. Rev.* **B46** (1992) 5621.
21. M. Greiter, F. Wilczek, and Z. Zhou, paper in preparation.
22. C. Nayak and F. Wilczek, *Nucl. Phys.* **B417** (1994) 359; *Renormalization Group Approach to Low Temperature Properties of a Non-Fermi Liquid Metal*, Princeton-IAS preprint, to appear in *Nuclear Physics.* **B**; H.J Kwon, A. Houghton, and J.B. Marston, *Gauge Interactions and Bosonized Fermi Liquids*, Brown preprint (unpublished, 1994); B.L. Altschuler, L.B. Ioffe, and A.J. Millis, *On the Low Energy Properties of Fermions with Singular Interactions*, MIT-Rutgers-Bell Labs preprint (unpublished, 1994); Y.B. Kim, A. Furasaki, X.-G.Wen, and P.A. Lee, *Gauge-Invariant Response Functions of Fermions Coupled to a Gauge Field*, MIT preprint (unpublished, 1994).
23. B. Halperin, P. Lee, and N. Read, *Phys. Rev.* **B47** (1993) 7312.

Persistent Currents in Normal Metals

Richard Webb

I.B.M.
Thomas J. Watson Research Ce ter
P. O. Box 218
Yorktown Heights, NY 10598, USA*

Abstract#

Evidence is presented for the existence of persistent currents in normal metals. It is shown that even in the mesoscopic domain, quantum effects may be very important. Investigations of the magnetic properties of metals in this domain have shown Ahavonov-Bohm effects that suggest that persistence currents should exist in normal metals. It is shown that experimental verification of the existence or non-existence of these currents is very difficult and not resolved at this time.

When you're in the laboratory trying to measure a quantum effect, you are often faced with many problems that theory may not have addressed. One interesting property is the possibility of persistent currents in normal metals. Since the technology is extremely advanced, no one can do these experiments without help from a large number of people. I wish to thank all those who have contributed to the efforts that made the results discussed here possible.

Theorists must understand that experimentalists can be very helpful. Just tell the experimentalists, in a way that we can understand, what it is you'd like to know. For example, consider a condensed matter system of some really macroscopic size and ask how to calculate the magnetism and the transport properties. We all know from classical physics how to do that. Then take the thermodynamic limit, do ensemble averaging over all possible scattering sites because, there are many of them there, and calculate an average magnetization or susceptibility or electrical resistivity for that material. But you know if you were to examine some small subsection of that sample, say a cube of atoms three on a side, twenty-seven atoms total, and ask what the magnetization or the transport properties are of that, your classical approach should break down simply because the electron's a wave, not a billiard ball. You' would have to invoke quantum mechanics. Now most experiments, until recently, could not get down to that kind of size scale.

The main discovery about six or seven years ago is that you don't have to go to the very small scale to see the quantum effects. There is another intermediate range, called the mesoscopic range by some, where the long range space coherence and the wave function provides you with ample quantum mechanical sensitivity to study. There is a correlated behavior over a length scale associated with the system. This range, called the space coherence length, is the distance an electron can move in a condensed matter system without losing the phase of its wave function. That distance, surprisingly, can be as long as twenty microns At IBM we are producing circuits of the future that are at the tenth micron level. They're going to be in your computers someday soon. So we're talking about systems that show extreme quantum effects, when cooled to low enough temperature, that are hundreds of times larger than the micro-circuits that we're building today. This is very exciting. Yet, they contain billions and billions of electrons, so we're really not dealing with microscopic systems.

Three years ago I reported on the state of research at that time. What was reported then is now an old story. At that time we used the then current state-of-the-art lithography to build a metallic system. We made a lithographic ring of gold about 1.86 microns in diameter, and did a four-terminal electrical resistance measurement. This ring was gold evaporated out of a relatively crummy, non-state of the art evaporator, using state of art lithography at that time.

When that system was cooled to low temperatures the behavior surprised many people in the community. What was discovered was that the electrical resistance that you measure as a function of the magnetic field oscillated periodically as the field was varied over 0.1 to 0.2 Tesla. That was a clear manifestation in some minds of an Aharonov-Bohm effect, and indeed that seems to be the most reasonable explanation.

There is another surprising feature that shows up if we use a half ring. If you look on a larger magnetic field scale, there additional fluctuations. The oscillations previously discussed occurred as the field was changed over a fraction of a Tesla. The new oscillations in the electrical resistance become apparent in the range of 0 to 8 T. These oscillations are weak and just visible on top of the previous oscillations if we use a complete ring. If we break the ring and only study one-half, we only see the new fluctuation effects.

Standard solid state physics textbooks say the electrical resistance of a piece of gold as the function of magnetic field is a smooth curve. That is what you should be teaching your graduate students. For years people thought these oscillations was a junk effect; however, when the theorist and experimentalist finally learned to talk to each other, what we finally understood was this also an Aharonov-Bohm effect. We can see that this is an Aharonov-Bohm effect by asking what would happen in a disordered system.

There are an infinite number of paths that the electron might take, but if we look at the intersection of any two paths that form a closed loop, and apply a magnetic field, the local probability of finding the electron at the intersection is a fluctuating function of magnetic field. This is caused by the interference phenomena associated with flux through that path. Obviously there are paths enclosing a wide variety of areas going from very small to essentially the sample size. So what you see is what many theorists have called an electron inferrogram. The conductance as a function of magnetic field fluctuates periodically. It's very reproducible, you can make measurements a month later and get the same pattern.

If you make another identical sample, you'll have a completely different pattern. The only universal thing is the amplitude. The amplitude is on the order of the electric field squared over the magnetic filed. In these experiments, the sample size has to be of the order or smaller than this characteristic phase coherence. The nice thing about these universal conductance fluctuations is we can measure the phase coherence length relatively accurately. Using a bit of theory, take an auto correlation function, then the half width at half-maximum is a measure of the phase coherence length. For a typical sample it is about 1.3 microns.

This is still an old story. We've been changing our experiments and asking new questions. What about the magnetic properties of the small system? I've always had a very small problem, which is that in our textbooks we teach that the magnetism of metallic systems is a combination of polyparamagnetism, which describes the coupling of electron spin to the applied magnetic field, and Landau diamagnetism, which describes the coupling of the orbital motion to a magnetic field. In this regime all the electrons are phase coherent, also what does the Landau diamagnetism do as a function of field in a phase coherent regime? Is it a number, or is it a fluctuating quantity that might have some Aharonov-Bohm effect? That's a question which we spent quite a lot of time trying to answer. A slightly different version of that is if you build a ring, then you're supposed to get persistent currents.

The basic idea is that there will be a current started as you put on a gauge flux to the center of a metallic ring. The characteristic current circulating around that ring will be an oscillatory function of magnetic field, or flux threading the ring. The period of oscillation will be Planck's constant divided by the electric charge ($\hbar$/e.

The theory is simple. Write the Hamiltonian for that system and consider the energy of each electron in each level to calculate the current carried by each of those states. This is just the derivative of the energy with respect to the flux. The magnitude of the current is the electric charge times the velocity in that state divided by the path. In an ordinary condensed matter system, the electrons occupy different levels. So if we have 10^{12} or 10^{13} the electrons, as we add them, are going to go into different Eigenstates. The very last electron added goes into the Fermi energy.

In a closed system of an annulus, all of the extensive properties are going to be highly periodic in the gauge flux. In particular, the energy levels are going to show a 2π periodicity for each flux level, the energy of the first electron will be an oscillatory function as will the energy of the next electron. In principle, the total energy oscillates periodically in flux, but this oscillation changes slope for each electron that you add. To first order, each new electron cancels the previous electron so as you work your way up this ladder, almost every contribution to the current is canceled by the one below it until you get to the Fermi energy.

In principle the persistent current you will get in this one-dimensional system is only eV_f/ℓ with eV_f being the Fermi velocity corresponding to the last electron you added, and therefore, its sign,. and ℓ is the coherence length. The response in a magnetic field can either be positive or negative. It only depends upon whether there is an odd number of electrons or an even number of electrons in your sample.

This is an Aharonov-Bohm effect. To this audience, that's probably not surprising, but most audiences believe that this is just Landau diamagnetism. What does this have to do with the Aharonov-Bohm effect? The Aharonov-Bohm effect is in the transport measurement. You send an electron in at some energy, it has two ways to get around the system, a displacement advances the phase along one path differently than along another path. When you re-combine the waves, you can get a phase difference, which gives you pattern of constructive and destructive energy.

You can analyze it another way. Break a ring into two parts. In part 1, the electron takes path. we denote by (1). Along this path the phase change is $\varphi_1 = \int_1 A \bullet dl$. Along the other part we indicate the path by (2) and the phase change is $\varphi_2 = \int_2 A \bullet dl$. To get the phase difference, subtract those two phases. One of the paths is oppositely directed, so we must add the phases.

To show there is a persistent current, draw an imaginary dividing line. Start an electron at that line and say it's going to go all the way around the ring, but divide it up into two separate paths, and then if you sum that up in terms of an Aharonov-Bohm effect, you'll see that the total change in phase is $\varphi_1 + \varphi_2$. Take a piece of gold that can be broken open and measure its electrical resistance. Now put it in a loop and it carries a persistent current. A persistent current to the age of the universe, not for a nano-second or a pico-second.

In the real world, one-dimensional rings cannot be built, at least not that have low disorder. We really have a multi-dimensional system where the line is 300-1,000 Angstroms thick. That is, lines are many electrons thick and many electrons wide. You might expect that there would be a multi-channel effect that would enhance the current. In the systems we have built, the electron scatters many times as it goes around the ring. It may scatter hundreds and hundreds of time before making one complete revolution. In that case the transit time for an electron to move around is basically a diffusive time as opposed to a ballistic time. The current is, in principle, thought to actually decrease because of the slow transit; whereas, in the ballistic case, if it doesn't scatter at all, the total current should be enhanced by the number of independent channels that you're carrying. So, in a real system you might expect some very large currents.

Experimentalists have learned many things in the last twelve years in condensed matter physics about the space coherence lengths. What should be obvious, but may not be to some, is that the characteristic distance which the electron moves without losing the phase information in its wave function can be shorted by the electron-phonon interaction, the electron-electron interactions, and the interaction of the electron with any magnetic impurities. Another criteria which all experimentalists have to be aware of is that there is a broadening of the energy of the wave packet due to a thermal diffusion process. We call this a dephasing length. This dephasing is due to a finite temperature effect. So you have to have this characteristic dephasing length along the perimeter of your ring. All this translates into typical ring sizes that are going to be on the order of 1 — 10 microns, and temperatures which must be milli-Kelvin.

We need to use a state of the art, or very close to state of the art, SQUID detection system. SQUIDs are just very sensitive detectors of magnetic field. They consist of a superconducting circuit which surrounds the ring under study. We apply an external magnetic flux, and if there's a signal, the signal will be coupled directly into the SQUID. Skipping the engineering details, you make this circuit such that it's a gradiometer where you wind two identical coils but in opposition to each other. Then if you apply a uniform field, without a sample in your SQUID coils, you'll get no signal coupled in your SQUID. Now the kind of sensitivity that I'll be talking about today, refers to the input terminals of the DC SQUID. We are able to resolve changes in flux to a part in 10^7 - 10^8. of a superconducting flux quantum. One-tenth to one-hundredth of a micro-flux quantum at low temperatures, with good signal averaging, is easily obtainable with these state of the art systems. Using modern lithography, we can make rings whose dimensions are on the order of 1-5 μm.

If we put a gold ring inside a niobium pick-up quill that's part of a SQUID circuit, how big should the persistent current be? Again we were surprised. This is another one of the things that should make you extremely doubtful about the existence of persistent current. If the calculation is done at temperature $\cong 0$ K using all the simple theory that's been out there for about five years, theory predicts a persistent current of 2.2×10^{-7} A/ℓ where ℓ is the coherence length in microns. In a one micron perimeter ring, the size of the persistent current should be 2×10^{-7} A.

A well equipped laboratory can routinely measure 10^{-15} A in transport experiments. This can easily be done with room temperature electronics and a little signal averaging. So it would seem that we can easily measure the persistent current in the ring. It is eight orders of magnitude bigger than what I normally measure. It turns out not be so easy after all; this is not an easy experiment. To see this, calculate the coupling of that ring to the detection system. The mutual inductance of that ring is about a pH. Now 10^{-7} A times a mutual inductance of 1 pH gives about 10^{-19} Volt-sec. That's about 10^{-4} to 10^{-5} of a superconducting flux. It's a small signal, even though the magnitude of the current is large.

What we actually do in the experiments is build these fantastic, highly versatile SQUID detectors. We then put many different samples, at many different locations in these detectors. For example, we used a gold ring that's about 1.4 by 2.6 microns, square as opposed to round. Cool these samples to low temperatures and collect the data. Ideally, if there was no signal coupled in, the measured magnetization as a function of applied field would be a flat line since the magnetometer is working in a balanced mode. Lithography is not perfect, so there is some imbalance which can be experimentally removed from the data. The data we obtain is a fairly straight-looking pattern, but the second order of correction looks like a cubic. That's just the response of the environment. There is no signal on top of that.

Simultaneously, while sweeping the DC field, we use AC techniques. As those of you who are experimentalists will know, you can get much better signal to noise by using phase sensitive detection. What we do is apply an AC Field and detect the AC response. We can use this to measure the fundamental response, or the next harmonic. Applying the AC field gives the primary, the second harmonic and the third harmonic response. That' is done simultaneously in this experiment, so we can get three of them at one time.

Do a little bit of signal averaging and background subtraction to get the fundamental signal. Subtract out a quadratic and what's left over is the reduced data. Fourier transform that to get a signal that is exactly like you would expect: an oscillatory signal based on the inside and outside diameters of that ring. This happens to be true for all our samples. If there is an h/2e signal, which I call in these experiments a higher order harmonic, it should be down by roughly 1/exp in these experiments. It is at least that low.

There's another bump in the reduced data that everybody points to. Anybody familiar with digital signal processing, that is a data on a finite interval, knows that once you subtract out a quadratic and a linear and a constant, you force the power in the Fourier transform to go to zero because you've subtracted it out. Then when you ask your computer to fit Fourier components to this, false bumps show up. This is really the tail of a signal from a very high frequency, most of whose power we have subtracted out. This is the result of data on a finite interval. It's instrumental; it has nothing to do with physics. If I could take field data over a bigger field scale, I would push this intensity towards zero.

If you then take a look at the second harmonic, by subtracting out only the linear part, this is the kind of data you get. That is about as good as one can get for oscillatory work. To get a better signal, use a bigger ring, but the signal dies quickly as the temperature increases so there is not much to study.

To try and prove there is a persistent current, we first study the ring and collect our data. We then warm the ring and etch the gold out of the ring. That is we just get rid of the gold, leaving everything else untouched. Re-cool the system and look at the size of the signal in the region where we found the *h/e* signal as a function of temperature. So we have the data for an empty magnetometer and the data for a filled magnetometer. There is a difference, so it looks as if the signal is real.

To get a good signal to noise, over a sweep from plus to minus thirty gauss, takes twelve to twenty-four hours. The experiment can be stopped at any point, held and it doesn't decay. The signal is persistent in that sense. We have an oscillatory magnetization whose average value, which I detect over long time scales, is unchanging. This hasn't been done for the age of the universe, nor for 10^7 seconds, but over relatively long laboratory time scales, it is constant.

The theorists want to compare this to the theoretical result. Theory would say that if you assume some simple exponential dependence on the basis of the thermal diffusion smearing the wave packet, it should be possible to account for the new ballistic system, the diffusive system and calculate the amplitude of the *h/e* signal.

The experiments at 5 milli-Kelvin have a signal that's about two orders of magnitude larger than theory. We're not measuring 10^{-7}A but we are in the 10^{-8}A regime. We find that the diffusive correction is not there. We can also find the sign of the effect; that is, determine if the response is paramagnetic or diamagnetic. A lot of people, still believe that this is a diamagnetic phenomena, they would expect the signal to be diamagnetic. Theory says 50% of the sample should be diamagnetic and 50% should be paramagnetic in the response to near zero field. Our two successful experiments so far indicate that the effect is paramagnetic. Recently, theorists have been working on this, and believe this discrepancy is due to an electron-electron interaction. I'm not going to go into that. However, electron-electron interaction does not seem to be a likely explanation since conventional theory does not allow any mechanism by which you can explain discrepancies which are a factor of 100.

Before we had published our work, Laurent Levy, Kerry Dolan Dunsmere, and Ellen Gushiah published a paper in which they too were interested in persistent currents. They had built what I call a hammer type sample, it's ten million copper rings. Each ring is about 0.5 microns by 0.5 microns on a side. There are 10 million of them, so if you can't measure one with one ring, maybe you can measure 10 million, Well, I set you up to believe that if there's a signal in this, it's going to be in an Aharonov-Bohm effect, *h/e*. And that's what every theorist thought. But when they published their data, they got a different signal.

Both the second and the third harmonic response functions ought to be oscillatory in the magnetic field. The phase relation here is zero, so it would just be 0° different. The second harmonic should be anti-symmetric about zero, so it should be zero at zero field. The third harmonic and the first harmonic should be maximum or minimum at zero field. When they analyzed their data and extrapolated to T= 0, they obtained an unexpected result. The signal was periodic, not an *h/e*, but in *h/2e*. The size of their current corresponded to 3.6 × 10^{-10}A per ring.

Is that an Aharonov-Bohm effect? Well, I don't know the answer to that. There's been a lot of words said that this *h/2e* effect is just a higher order effect. In fact, I've been worried about some of the interpretations. What we've been doing lately is studying arrays of gold rings, but now we've developed a better detection system, I don't need 10 million, only an array of 200. Using lithography, the experimentalists can actually tailor the detection system so that each pick-up quill fits around one ring. Make the pick-up quills as small as needed, or the rings as big as needed, and use lithography to determine the optimum detection configuration. The beauty of our experiments is that we simultaneously can apply a magnetic field to both sides of the sample. The other side doesn't have any rings in it, so you can get rather uniform fields over the sample. Then vary the magnetic field continuously. We find a peak right in the vicinity of *h/2e* and some structure in the vicinity of *h/e*. However, the *h/2e* dominates.

In going from one ring to 200, there is a phenomena that's occurring, something is bigger. In our original single ring experiments, we could determine the $h/2e$ effect was about 1/exp down from the h/e effect. But, it's only true that on average. Every ring has a different number of electrons, so the h/e effect is only going to grow like the square root of the number of sample. If there's something more correlated, like an $h/2e$ effect, the correlated signal is going to grow like the number of samples. For 200 samples you expect the h/e signal be about 10 times larger, and $h/2e$ signal to be 200 times larger than that for a single ring. Those are the experimental results, roughly. We find the h/e signal is smaller than theory, and the $h/2e$ signal is about what is expected.

Although Levy's experiment was originally published as in perfect agreement with theory, it's now generally recognized that signal that he was measuring is about one to two orders of magnitude larger than they should be measuring based on the current theory. Our experiments also give the 1-2 order of magnitude difference. So we have two independent experiments both giving something much larger than they should.

Recently A. Benoit and his colleagues at CNRS-Grenoble have been studying the persistent current in a single gallium arsenide ring. This is a beautiful experiment (unpublished at the time of this lecture). This is where the foundations of quantum mechanics is going to really learn something. He, first of all, builds a ring with four terminals out of gallium arsenide in the ballistic regime, so the electron has no scattering. Then he can measure the h/e oscillations and Fourier transforms them to get the power spectrum. The electrical resistance oscillates periodically in both an h/e and about 1/exp down, on $h/2e$ component.

This is in the transport, no new news here. But now this is gallium arsenide, so he can put gates on top of it and deplete the electrons from the leads,. He then isolates it. and builds around the same ring a DC SQUID system, and now measures the magnetization of that isolated ring. Using the DC SQUID he finds a current going around in the isolated ring. He sees an h/e signal. The signal to noise is weak, but none of these experiments have good signal to noise. The beauty of this experiment is that now you can couple the ring to the outside world by taking the voltage off the gates. When that is done, this signal goes away. I think there's something significant there for the foundations of quantum mechanics and the whole idea of measurement theory.

That's about all I have to say. I just wanted to sort of summarize by saying that the micro-electricity industry is now providing samples where we can start testing some of the more fundamental predictions of quantum mechanics.

References

* Current Address: Department of Physics, University of Maryland, College Park, MD 20742, USA

# The material in this paper is a summary from the video tape of the talk presented by Dr. Webb at the meeting. Any errors in interpretation are those of the editors and should not be blamed on the author.

L. Levy , et al., PRL E4, 2074 (1990)

SECTION 4

BLACK HOLES AND QUANTUM GRAVITY

EVIDENCE FOR A MASSIVE BLACK HOLE IN THE CENTER OF OUR GALAXY

CHARLES H. TOWNES

University of California, Physics Department
Berkeley, California 94720, U.S.A.

ABSTRACT

Use of wavelengths other than the visible have recently allowed astronomers to study the center of our own galaxy, until now hidden by interstellar clouds. The densities, ionization states, temperatures and velocities of gases and dust near the galactic center tell us the radiant energy present and that the gravitational field corresponds to a black hole of 2-3 million solar masses at the center. More recent measurements of stellar velocities in the region confirm this evidence. However, precise identification of which object may correspond to a massive black hole and explanation of other phenomena observed in the galactic center are still matters of debate.

Our galaxy has many massive dark clouds composed of common molecules and dust. So many clouds lie between us and the center of our galaxy that we obtain no detectable visible light from the galactic center, and hence until rather recently astronomers were not able to study this important region. During the last few decades the use of radioastronomy, improving technology in the infrared region, and the availability of spacecraft to measure x-rays and gamma rays have all given us opportunities to detect radiation from this region and as a result we now know much about it, even though there are still puzzles.

High resolution radioastronomy has identified a rather powerful point source[1] of continuum radiation rather close to the dynamic center called Sgr A*. In addition, there is an oval shaped ring of fast moving ionized gas[2,3,4] corresponding to the projection of a circular ring rotating at approximately constant velocity around Sgr A* and at a distance of about 4 ½ light years. Outside the ring are molecular clouds of varying density but generally in the range of 10^4 to 10^5 molecules per cubic centimeter. Inside the ring there are blobs of ionized gas of similar density, also regions in which almost no gas exists, and at least one sizeable atomic gas cloud. Analysis of the velocities of these gases indicates that inside of the ring there must be a total of about 4 ½ million solar masses and that there must also be a concentrated mass in the center of a few million solar masses[5]. Overall, both the ionized and the molecular clouds are not in a steady state configuration or velocity distribution, indicating that within the last hundred thousand years some rather violent phenomenon must have taken place. This might have been several very large supernova explosions, though the total magnitude of the disturbance is almost too large to explain this way.

Recent high sensitivity and high resolution infrared cameras have been able to detect a number of hot stars in the central region[6], with concentrations especially high within about 1 light year of the radio point source Sgr A*. In addition, the velocities of cooler stars in this region have been measured from the spectrum of CO in their atmospheres[7]. It is found that these velocities correspond rather well to velocities of the gas already mentioned. However, the stars give a somewhat more secure measurement of the velocity distribution and hence the gravitational field in the region than does the gas, because there has always been an uneasy feeling that some other mechanism might possibly have accelerated the gas, such as varying magnetic fields. In fact, however, measurement of the magnetic fields through Zeeman effects[8] on atomic and molecular transitions indicate that the fields are less than about 1 milligauss and too small to have very much effect on the dynamics of the gas.

General expectations for the source of mass in the center of the galaxy have been that it would be either a dense collection of stars broadly similar to the globular clusters which are very familiar to astronomers or that there might be some combination of stars and a central black hole due to material continually falling into this gravitational well. If the mass is due to a cluster of stars alone, then because of interstellar collisions the stars would on the average have the same velocity independent of the distance from the center and a density distribution proportional to $1/R^2$, where R is the distance from the center. On the other hand, if the gravitational field is produced by a single point mass or black hole, the velocities of stars or gas would be proportional to $1/R^{1/2}$. In fact, at distances greater than about 5 light years the velocities appear to be dominated by stars and are constant as a function of distance from the center. Inside of this distance, however, there are deviations from constancy. The increased velocity with decreasing distance is particularly noticeable inside of a few light years and indicates the presence of a very concentrated mass at the center of 2 or 3 million solar masses. The only form which theory presently allows for such a concentration is a black hole.

Although Sgr A* is a good candidate for a black hole, nevertheless neither it nor any other object near the center is presently producing the spectacular phenomena we normally expect from a black hole into which much material is falling. Long baseline radio interferometry has been able to demonstrate that Sgr A* is quite stationary or moving only very slowly, at velocities less than about 25 km per sec[9]. Other objects in the same region characteristically move at least about 200 km per sec. Hence, there is evidence that Sgr A* must be substantially more massive than other stars or objects in the same region. While this source emits radio waves and infrared with characteristics somewhat like those expected from a black hole, the total radiation at the moment is quite weak compared with normal expectations. Perhaps material previously falling into the black hole produced such a violent generation of energy that materials have been blown away in the recent past, perhaps with the event which must have disturbed the clouds during the last hundred thousand years and blown gas away from the center. However, there is presently some gas close to the source and we must suppose that either the generation of energy is unusually low at this particular moment or that this black hole is behaving somewhat differently from our expectations.

Observations of x-ray radiation from satellites enlarge the puzzle of Sgr A*. While there are some x-rays coming from the region of the galactic center, they are relatively weak. Furthermore, because x-rays would be scattered by clouds surrounding the galactic center, one can look for the scattering and hence trace something of a history of any powerful production of x-rays from the center over the last few thousand years. Some of the x-rays would have moved out into our galaxy a few thousand light years and then been scattered towards us. Evidence shows that Sgr A* was a relatively weak source of x-rays even throughout the last few thousand years[10]. In spite of this lack of production of the high power which is normally expected of a black hole, the gravitational evidence based on velocities of gases and stars seems to provide rather clear evidence for a high concentration of mass, presumably a black hole. Furthermore, the gravitational field of a black hole is a characteristic about which we cannot be mistaken, whereas the generation of energy from infall represents a much more complicated theoretical problem, which faces us with some uncertainties.

Characteristics of the ionized gas and warm dust radiation from the central few light years of the Galaxy indicate the presence of intense ultraviolet radiation and a total luminosity about 10^7 times that of the sun. These characteristics appear to be explainable by the presence of a few tens of rather hot (T ≈ 30,000 K) stars in this region which have recently been detected. Why these stars are present, however, is a puzzle. If there was star formation from gases near the center, it must have occurred within the last few million years and have formed a very unusual collection of stars. Furthermore, present conditions in the galactic center do not seem favorable for star formation. Perhaps instead, these stars represent mergers of several stars in this region of high stellar densities, somewhat as the "blue stragglers" in globular clusters are thought to have been formed. At present, their formation and character are puzzling, as is also the exact nature of the unique source, Sgr A*.

The great progress recently made in observations of the galactic center have been due to important technical and instrumental developments as well as vigorous astrophysical research. Fortunately, we can expect further instrumental progress and hence perhaps a thorough understanding of the very interesting laboratory which is our galactic center, and the remarkable phenomena occurring there.

References

1. B. Balick and R.L. Brown, *Ap.J.*, 194, (1974) **265**
2. K.Y. Lo and M.J. Claussen, *Nature*, 306, (1983) **647**
3. J.H. Lacy, C.H. Townes, T.R. Geballe, and D.J. Hollenbach, *Ap.J.*, 262, (1982) **120**
4. E. Serabyn and J.H. Lacy, *Ap.J.*, 293, (1985) **445**
5. R. Genzel and C.H. Townes, *Annual Review Astronomy and Astrophysics*, 25, (1987) **377**

6. A. Krabbe, R. Genzel, S. Drapatz, and V. Rolacive, *Ap.J. Lett.*, 383, (1991) **19)**
7. K. Sellgren, M.T. McGinn, E.E. Becklin, and D.N.B. Hall, *Ap.J.*, 359, (1990) **112**
8. N.E.B. Killeen, K.Y. Lo, and R. Crutcher, *Ap.J.*, 385, (1992) **585**
9. D.C. Backer and R.A. Sramek, *URSI meeting*, Boulder, Colorado (1993)
10. R. Sunyaev, to be published

BLACK HOLES, WORMHOLES, AND THE DISAPPEARANCE OF GLOBAL CHARGE[†]

SIDNEY COLEMAN
Department of Physics, Harvard University
Cambridge, MA 02138 USA

ABSTRACT

One of the paradoxes associated with the theory of the formation and subsequent Hawking evaporation of a black hole is the disappearance of conserved global charges. It has long been known that metric fluctuations at short distances (wormholes) violate global-charge conservation; if global charges are apparently conserved at ordinary energies, it is only because wormhole-induced global-charge-violating terms in the low-energy effective Lagrangian are suppressed by large mass denominators. However, such suppressed interactions can become important at the high energy densities inside a collapsing star. We analyze this effect for a simple model of the black-hole singularity. (Our analysis is totally independent of any detailed theory of wormhole dynamics; in particular it does not depend on the wormhole theory of the vanishing of the cosmological constant.) We find that in general all charge is extinguished before the infalling matter crosses the singularity. No global charge appears in the outgoing Hawking radiation because it has all gone down the wormholes.

[†] This abstract is a report on work done with my graduate student Shane Hughes. A full description of the work is available in *Phys. Lett.* **B309**, 246 (1993).

THE CONFLICT BETWEEN QUANTUM MECHANICS AND GENERAL RELATIVITY

LEONARD SUSSKIND
Department of Physics, Stanford University,
Stanford, California 94305-4060

It is a great pleasure for me to contribute to Yakir Aharonov's festschrift. Over the past three decades that we have been close friends, I, like so many others, have found Yakir's profound insights truly inspirational. The only subject that I can remember us disagreeing about is the quantum mechanics of black holes. It is a small irony that I choose this topic for Yakir's celebration.

Introduction

It is almost one hundred years since the discoveries of the quantum and of special relativity. It has taken most of the twentieth century to synthesize these into the modern quantum theory of fields and the standard model of particle physics. By contrast almost nothing is known about the connection between quantum mechanics and the general theory of relativity. The relevant phenomenon are too remote and inaccessible to experiment for us to expect much guidance from that direction in the foreseeable future. For this reason most work on the subject has been guided by purely mathematical considerations.

I believe that we need more than this to keep us on the path of phenomenology (what used to be called physics) and not wild speculation, and that in the absence of real experiment our only hope is to focus on gedanken experiments involving realistic situations which may be beyond our technological capabilities but are otherwise possible. Perhaps we will uncover physical paradoxes and puzzles whose unraveling will provide deeper insight than we now have. Let me just remind you how much was learned from the paradoxes concerning the constancy of the speed of light, the finiteness of specific heat of radiation and the stability of the atom.

Why then black holes? The reason is a combination of factors. First of all black holes are real objects which can be assembled from ordinary matter. To think that black holes can not exist or have never formed is far more radical than to assume the opposite.

Secondly, we know from the work of Bekenstein and Hawking that black holes are catalysts from new phenomena that intimately involve gravity and quantum mechanics. Magnetic monopoles also act as catalysts of, otherwise, very remote phenomena, namely the violation of baryon conservation. In the case of black holes, we do not know with certainly what the catalyzed effects are but there is reason to believe that they are more far reaching and profound than baryon violation. They may even involve the breakdown of the principles of quantum mechanics.

The central problems I will discuss has been with us since Hawking's remarkable observation that black holes evaporate. The main reason it has stimulated recent interest is the discovery of 1+1 dimensional theories containing black holes. Initially it was thought that these theories were so simple that surely we could completely analyze them and discover the precise nature of black hole evaporation. This is not what has happened. The 1+1 dimensional theories have just reinforced Hawkins original arguments leading to his disturbing conclusion that black holes seem to catalyze a breakdown of quantum mechanics.

Black Holes and Thermodynamics

In 1973 Bekenstein raised the question of whether the second law of thermodynamics could be violated by dropping thermally excited matter into a black hole so that its entropy could be caused to disappear. Based on Hawking's observation that the total area of black hole horizons always increases, Bekenstein postulated that a black hole has an intrinsic entropy proportional to its area measured in plank units. The precise formula is

$$S = \frac{\text{area}}{4}.$$

Since the mass and area are related by

$$A = 4\pi R^2 = 16\pi M^2,$$

one has a connection between entropy and energy

$$S = 4\pi M^2$$

If one also postulates the usual thermodynamic relation

$$dE = TdS.$$

Then the temperature of a black hole is

$$T = \frac{1}{8\pi M}.$$

That a black hole should have entropy is not so surprising. Entropy is a measure of ignorance. More exactly it is the logarithm of the number of macroscopically indistinguishable microstates of a system. Since from the outside one can never tell what a given black hole was formed out of, it is reasonable that it has an entropy. It was more surprising that it has a temperature.

Hawking soon realized that the finite temperature should cause a black hole to radiate like a black body. Indeed, Hawking was able to show by quantum field theoretic means that a black hole radiates like a body of area ~16π M^2 at exactly the temperature $\frac{1}{8\pi M}$. Thus its luminosity is given by the Stephan Boltzmann law

$$\frac{dM}{dt} = \text{Luminosity} \sim \text{area} \times T^4 \sim \frac{\text{const}}{M^2}$$

It therefore follows that the black hole radiates away its energy in a time ~M^3. The radiated energy is thermal with a gradually increasing temperature. A solar mass black hole would have a temperature ~10^{-11}ev which would make it far cooler that the ambient microwave background. It would therefore absorb radiation and grow. A million ton black hole would have a temperature of order 10 GEV and a lifetime ~10^9 sec.

Evaporation of the black hole is not in itself a problem. The paradox announced by Hawking concerns the fate of information which falls into the hole. Let us consider two distinct (orthogonal) ways of producing a black hole of a given mass. The two configurations may be a collapsing neutron star, the other an antineutron star. The difference might be more subtle, consisting of only a single neutron being replaced by an antineutron. In either case the two initial configurations are described by orthogonal vectors. How many distinctly orthogonal configurations can produce a black hole of mass M? Classically the answer is infinite. If however we are to believe the usual connection between entropy and information, the result should be

$$\sim \exp S \sim \exp M^2.$$

On the other hand, the no-hair theorem tells us that the geometry outside the horizon is unique. Hawking's calculation of black hole radiation only depends on the exterior and produces featureless thermal radiation which in no way depends on the details of the in falling matter which produced the black hole. Evidently this information is lost unless

1) The black hole ceases evaporating leaving a remnant containing the information.
2) A more complete computation of the Hawking radiation which includes the quantum dynamics of the horizon produces a mechanism for imprinting the information on the Hawking radiation.

An S Matrix?

'tHooft has phrased the question as follows: The initial state consists of a set of ingoing particles. The particles could be composites such as atoms, planets, Encyclopedia Brittanicas (for some reason theorists love to throw encyclopedias into black holes) etc. The outgoing stuff is also particles which in some approximation look like thermal radiation. If ordinary quantum mechanics describes the event of formation and evaporation then it must be described by a unitary S matrix.

$$S|in> = |out>.$$

Since the S matrix is unitary the initial state should be reconstructable from the final state

$$|out> = S^{+}|in>,$$

thus quantum mechanics forbids the erasing of information.

Let me be a little more precise. In general, quantum mechanics will not allow us to reconstruct an initial state by doing a set of experiments on the final products of a single event. I have in mind an ensemble of events, all prepared in identical manner. In some of these events I measure a complete commuting set of operators which provides a probability function in this basis. In another subset of events, I measure another set of operators which do not commute with the first. With enough such measurements the quantum state of the final radiation can be obtained. It should be a pure state.

Now do the whole procedure over with an initial state which is orthogonal to the first. The resulting final state should be orthogonal to the first. The problem is that according to Hawking's calculation the products of evaporation consist of absolutely thermal radiation.

So what, you say. Exactly the same thing happens when a bomb goes off. The initial detailed features of the bomb are erased but no one thinks quantum mechanics is violated. It is instructive to consider this even in more detail. Let's suppose the explosion takes place in a cavity with perfectly reflecting walls except for a small hole where radiation can leak out. The initial state consists of empty cavity plus bomb plus encyclopedia. After the explosion the cavity is filled with hot gas and radiation which soon comes to equilibrium. Radiation slowly leaks out. Eventually the box is in its zero-temperature ground state and the thermal entropy of the outside world is increased by the outgoing radiation.

I will begin analyzing this experiment by first considering two kinds of entropy which exist in quantum mechanics. The first I will call entropy of entanglement. Consider two subsystems, A and B. In our example, A is the region outside the cavity and B is the inside region. Assume the space of states is a product $H_A \otimes H_B$. A wave function is a function of the coordinates of a and those of b.

$$\Phi\ (a,b)$$

The density matrix of b subsystem is

$$\rho_b = \sum_a \Phi^*(a,b)\Phi(ab')$$

and that of the a subsystem is

$$\rho_a = \sum_b \Phi^*(a'b), \Phi(ab)$$

The entanglement entropy associate .vith a given density matrix is

$$S_E = -\mathrm{Tr}\, \rho \log \rho$$

Thus, in general, the subsystems A and B have entropy due to entanglement. Furthermore it is very easy to prove that

$$S_E\,(A) = S_E\,(B)$$

The only situat .n in which S_E is zero is Φ (a,b) is an uncorrelated product Φ (a) Φ (b). If th^ interaction between A and B is switched off S_E (A) and S_E (B) are conserved. The entanglement entropy is not the entropy of the second law. One final point is if the dimensionality of H_A is D_A then the maximum value S_E (A) (and therefore S_E (B)) can have is -log D_A.

The second kind of entropy S_I is thermodynamic entropy or entropy of ignorance. Sometimes we assign a density matrix to a system, not because it is quantum-entangled with a second subsystem, but because we are ignorant about its state. We assign a probability to different states. For example if we know nothing about a system, we assign the unit density matrix. If we know only the energy we assign a projection operator $\delta(E - E_0)$. In thermal equilibrium we know the probabilities for a small subsystem to have energy E and we assign the Maxwell Boltzmann density matrix.

$$\rho_{MB} = \frac{1}{Z} \exp\left(-\beta H\right)$$

The entropy of ignorance is always larger or equal to the entanglement entropy.

Now, following Don Page, let us consider the time dependencies of the various entropies in our experiment. Begin with the thermal entropy S_I (B). At first it's zero because we assume everything is know about the bomb-box system. Actually there may be a small entropy of entanglement with the outside but if the inside and outside are weakly coupled it is small. The bomb explodes and the thermal entropy suddenly increases to some maximum characterized by some initial temperature T. As time evolves, the box cools and the thermal entropy decreases to zero.

Now consider the thermal entropy outside the box. It begins at zero and gradually increase as the thermal radiation escapes. According to the second law, its final value exceeds the thermal entropy in the cavity just after explosion. Fig. 1 illustrates the evolution.

Now consider the entanglement entropy. Since they are equal inside and outside, we only need to consider the inside of the cavity. Since the cavity is initially almost unc rrelated with the outside S_E is zero. This is still true shortly after the explosion. However as photons leak out the inside and outside become entangled and S_E (B) increases. Eventually however since S_E (B) ≤ S_I (B) it tends to zero. This is because the cavity returns to the ground state. This evolution is shown in Fig. 2. Evidently the final exact density matrix of the outside is significantly different in some subtle respects from the coarse grained density matrix ascribed to it by the thermal description.

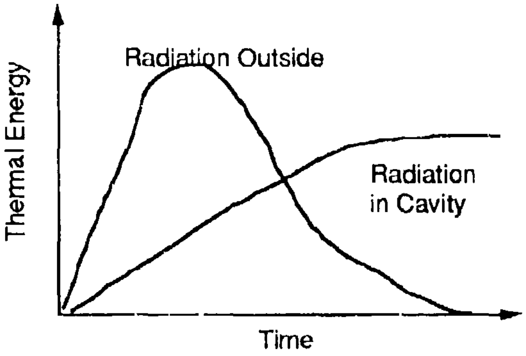


Figure 1: Evolution of Thermal Energy Inside and Outside a Box.

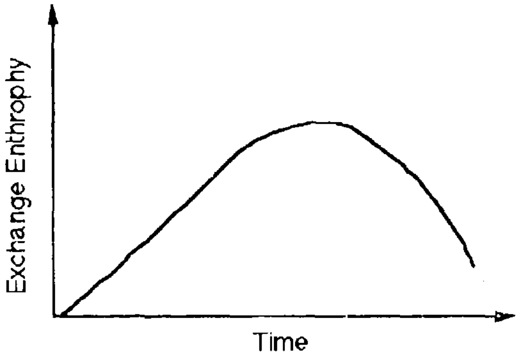


Figure 2: Evoluton of Entanglement Entrophy.

To understand the difference consider the time at which S_E is maximum. The entanglement entropy outside the box may be comparable to the thermal entropy. At this time large correlations exist between the outgoing radiation and cavity. Later on when the box has cooled, those correlations become correlations between the radiation which came out early and the lately radiated photons. In other words, the subtle way in which the outgoing radiation is not exactly thermal is the existence of long-time correlations. Locally the radiation looks extremely thermal. It is these correlations which carry all the initial information . The central question facing black hole theorists is whether such subtle long-time correlations exist in the Hawking radiation accompanying black hole evaporation. The dilemma is that if they do not, then the process of formation and evaporation cannot be described by an S-matrix and ordinary quantum mechanics can not describe it. However no known mechanism exists for transferring the information from the infalling matter to the outgoing radiation. Let us see why this is so.

Penrose Diagrams

A Penrose diagram is a schematic representation of a space time which is especially useful for spherically symmetric situations such as a Schwartzshild black hole. All of space time is represented on a finite region with time-like, and space-like infinities mapped to points. For example empty flat space time is shown in Fig. 3.

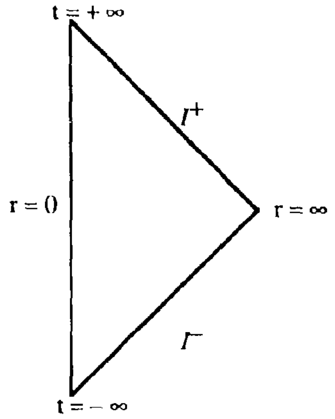


Figure 3: Penrose Diagram for Flat Space-time.

The lines labeled $I^\pm$ are called past and future light-like ∞. They are the places where light signals begin and end. All radial light-signals are represented by 45° lines.

An eternal black hole is shown in Fig. 4. The wavy dark lines are past and future singularities and the past and future horizons are dashed lines.

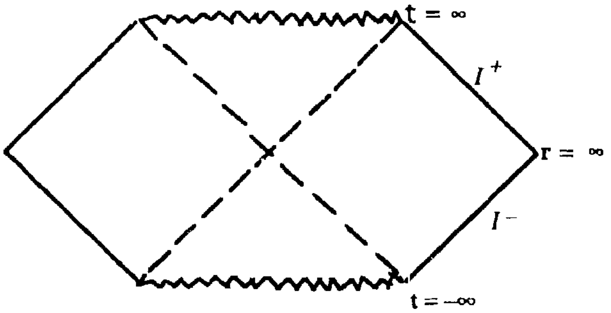


Figure 4: Penrose Diagram of an Eternal Black Hole.

In classical general relativity, a black hole can be formed from infalling matter but does not evaporate. The Penrose diagram for a black hole created by an infalling massless pulse of radiation is shown in Fig. 5.

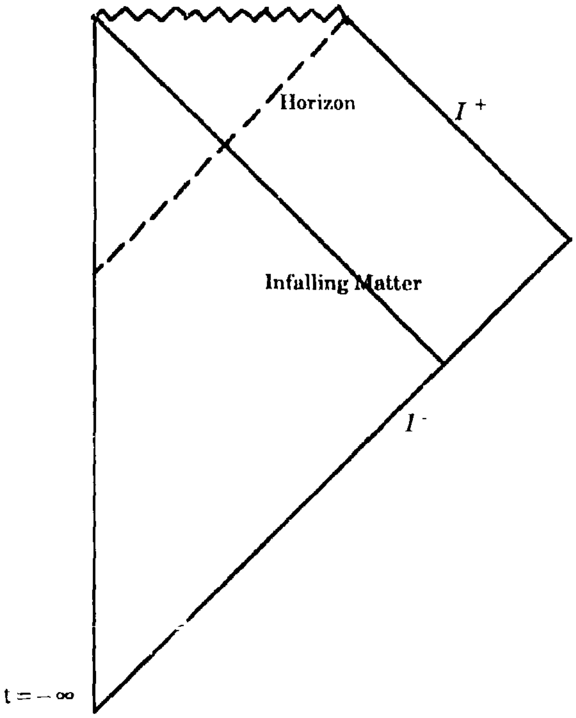


Figure 5: Penrose Diagram of a Black Hole formed by Infalling Matter.

Finally the Penrose diagram for the formation and subsequent evaporation of a black hole is shown in Fig. 6

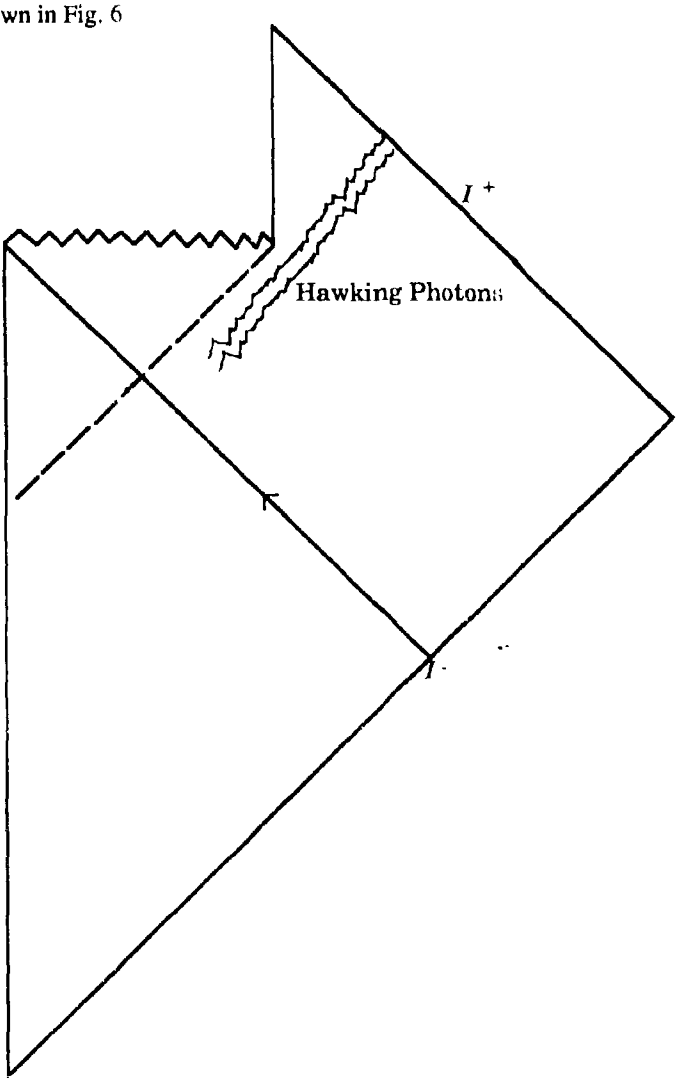


Figure 6: Penrose Diagram of a Black Hole that forms and then evaporates.

Let us now consider a spacelike surface Σ which consists of a part inside the horizon and a part outside as in Fig. 7.

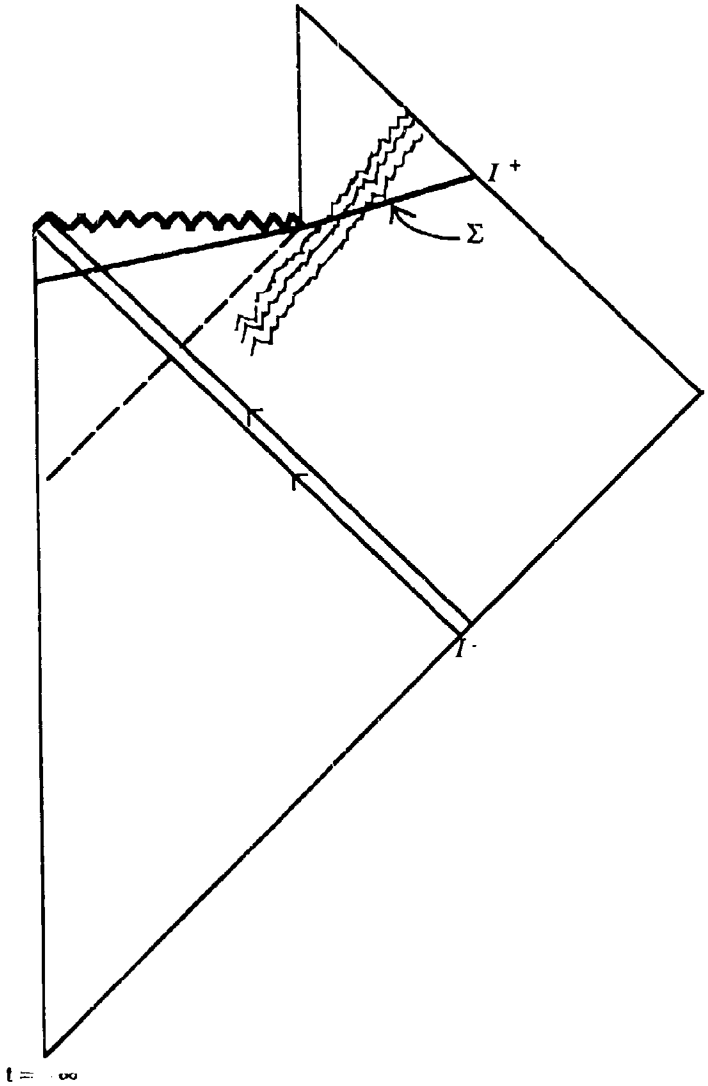


Figure 7: Spacelike Surface Σ across the Horizon .

The spacelike surface intersects both the infalling matter and the outgoing radiation. According to standard quantum field theory, we can specify a quantum state on this surface which lives in a product Hilbert space which by analogy with our previous discussion we label $H_A \otimes H_B$ where H_B is the state space inside the black hole. Since to the future of Σ the inside and outside evolve independently, their entanglement entropy are separately conserved. If the entropy of the radiation is to be zero, it must already be zero on Σ. This in turn would require the state on Σ to be a product, Φ (a, b) = Φ (a) ϕ (b). Let us also call the incoming state on I-, $|x >$ in. Let us denote the initial state dependence of Φ (a) and ϕ (b) by Φ_x (a), ϕ_x (b).

$$|x > \rightarrow \Phi(a)\phi_x(b)$$

Now assume that an outside observer who sees only A can describe such events by a unitary S matrix. This requires the observed final state Φ_x (a) to be linearly related to x. From the form of the final state Φ_x (a) ϕ_x (b) it follows that if Φ(a) is linear in x, ϕ (b) must be independent of x. The meaning of this conclusion is that there is no way that all of the information in x can escape in Hawking radiation unless it is completely obliterated before crossing the horizon. The obliteration of the initial state is however at odds with our usual conception of the horizon. It is almost universally believed (including by me) that an infalling observer feels nothing unusual as she crosses the horizon. Classically the horizon of a large black hole is locally very flat with no large deviations from the flat vacuum. Furthermore quantum field theory in such a background indicates no significant quantum corrections to the flat-space vacuum. All indications are that the information contained in infalling matter is not deflected as the horizon is approached. By an argument similar to the above, one ought to conclude that the state of the Hawking radiation is independent of the infalling matter so from an operational point of view, quantum mechanics would be violated.

The trouble with this viewpoint is that it does not illuminate the meaning of the Hawking entropy. If we think that entropy has its usual meaning then the entanglement entropy of the decaying black hole (or the radiation) should be less than or equal its thermal value which according to Hawking tends to zero as the mass evaporates. In fact if an extremely large black hole of mass M evaporates to mass m (still much bigger than Planck) the outgoing radiation should have thermal entropy larger than M^2 but entanglement entropy smaller than m^2. This should mean that long-time correlations carry out large amounts of information. Any other resolution of the information paradox should also explain why ordinary thermodynamics works for an outside observer without the usual underpinnings of standard quantum mechanics.

I have spoken with a large number of people from both the particle physics and the gravity communities, some of whom I consider very deep thinkers. I have found no clear pattern in their opinions. Hawking is strongly convinced that information loss is catalyzed by black holes while 'tHooft is equally convinced that an S matrix exists. Aharonov is the champion of a group who believe that plancksized remnants store all the initial information. John Wheeler would only say, "Hmmm, this is disturbing".

Perhaps that is all that should be said now. As for myself, I believe the foundation of quantum mechanics and information theory must be correct and that an S matrix exists. An outside observer sees the horizon as a thermally excited membrane. I believe that a correct description will be found in which, like all real membranes, the information stored on it when it is thermally excited will be accounted for as it returns to its ground state. Nevertheless, I also believe that a freely infalling observer sees nothing special at the horizon, but since he can not communicate this fact to the outside, no contradictory conclusions will be reached by an observer. However, at the present time this view seems inconsistent with the traditional ideas of local quantum field theory which would demand that the question of whether an infalling observer passes through the horizon or is disassembled into bits and radiate as Hawking radiation would have an invariant answer. Perhaps this is one of those times where progress can only be made by simultaneously believing two apparently inconsistent things.

UNITARITY OF THE BLACK HOLE SCATTERING MATRIX

G. 't HOOFT
Institute for Theoretical Physics, University of Utrecht, Postbox 80 006
3508 TA Utrecht, the Netherlands

ABSTRACT

Approaches towards the problem of constructing an *S*-matrix for a black hole are outlined. An earlier proposal by this author showed that this *S*-matrix will be related to string theory amplitudes (though they are not identical). A new approach formulated here involves the entire black hole history, for which a topologically trivial Penrose diagram is constructed. We can then construct segments ("blocks") of the *S*-matrix, for which there is no problem of information drainage. Because of this it is suspected (though not proven) that the *S* matrix constructed along such lines will be unitary.

1. Introduction

The problem of reconciling the theory of general relativity with the principles of quantum mechanics is one of the deepest and most fundamental ones of theoretical physics and it continues to mystify many of us. Now the procedure of replacing a "classical" theory by a corresponding "quantum mechanical" one is straightforward in many cases, in particular when we are dealing with relatively tiny interaction strengths or a small number of degrees of freedom. Indeed, if we consider circumstances where the gravitational force is weak and therefore accessible to a perturbative treatment we know fairly precisely how to perform this so-called "quantization procedure". The resulting theory, perturbative quantum gravity, turns out to be similar to any other gauge theory, except that when increasing accuracies are required new, undetermined physical parameters emerge: subtraction constants associated with unrenormalizable interactions. This complication, though of course a fundamental one, is relatively mild compared to the obstacles one encounters whenever a "non-perturbative" formalism is asked for. One then notices that any attempt even at giving a sensible frame for a description of what might happen will falter at distance scales smaller than the Planck length. A fundamentally new approach is needed.

One reason why any attempt based on the classical description of gravity must break down is a basic instability of the gravitational force: the possibility of gravitational collapse. As soon as too much energy is concentrated within one tiny volume element, a black hole - sometimes of considerable size - emerges. In a theory of quantized elementary particles something at least as complicated is expected to heppen. But this would then be a drastic deviation from one of the basic starting points in quantum field theory, namely that particles can be treated, in a first approximation, as if they move independently of each other, as if particles states can be simply superimposed on top of each other. In the high energy regime this must be utterly false.

If we add to this the observation that the high energy regime of a particle theory is connected to the low energy regime by Lorentz transformations we see that this brings into doubt either the basic postulates of Lorentz invariance, or the superposition principle for quantized particle fields, the applicability of partial differential equations to these fields, and the like.

Whenever extremely strong gravitational fields come into play we encounter fundamental problems of this sort in our understanding of the basic laws of physics. The strongest gravitational fields possible occur in the vicinity of black holes. This brings us quite naturally to the consideration that indeed black holes are the prototype testing facilities for any quantum gravity theory. A proper incorporation of black holes in any theory of quantized gravity must be absolutely essential, since they form the natural asymptotic limit of the energy spectrum of "most pointlike" particles.

And most standard theories of gravity do not incorporate black holes properly. In a proper theory black holes, or at least objects that would behave like black holes in the limit of large mass and size, should occupy a natural position in Hilbert space, be included in the unitarity conditions of the S martix, and so on.

In stead, what is usually done is that black holes are treated in the so-called background formalism. One specifies the metric as if it were a classical one, and then performs quantum field theory with respect to this background. At first sight one would expect that this were a correct procedure, comparable to, for instance, the treatment of magnetic monopoles in a gauge theory for elementary particles. But the outcome is drastically, and catastrophically, different[1]. It is found that, when viewed this way, quantum black holes extract and destroy "quantum information"[2]. In terms of pure quantum states this means that when we start with two states that are orthogonal to each other in Hilbert space, for instance because they differ by the presence of one extra particle moving into the black hole, these states become indistinguishable after a while, and hence cannot continue to be orthogonal to each other; if they did, the number of possible states inside a black hole would rapidly surpass the total number of possible states in the universe. In a slightly different interpretation of the same mental exercise one would say that a quantum mechanically pure state evolves naturally into a quantum mechanically mixed state.*

* A similar phenomenon seems to occur in theories of multiply connected universes. Here an uncertainty in the fundamental interactions arises on top and above the familiar quantum uncertainties. Pure

One could try to maintain, as indeed is often done, that black holes must therefore be radically different from elementary particles, including solitons such as magnetic monopoles. But this is too rash a conclusion. It would imply that black holes are not even "quantum predictable", but only obey probabilistic laws. To this author such a lack of precisely defined physical equations such as the Schrödinger equation is not likely. Surely the background metric approach to black holes cannot be right, just because it assumes that the particle fields can be superimposed onto the background fields, and we had already concluded that this superposition principle cannot be correct.

We can pinpoint in another way the complication that was ignored: there are interactions, in particular gravitational ones, between the in- and outgoing particles. Now under normal situations this would not have been a great disaster. In quantum field theories one can easily correct for such interactions by adding a series of successively tiny perturbative corrections. But the gravitational interactions are not normal in this respect. If we want to know how the out [illegible] react upon any variation among the ingoing particles at an earlier epoch, we find a disturbing divergence: the strength of the mutual gravitational interaction diverges *exponentially* with the time difference. Hence any perturbative approach is out of the question whenever we wish to follow the evolution of some configuration over any appreciable time interval.

In these notes I will skip the general introduction to black holes, which have been described abundantly in the literature[1-4]. One important aspect one has to remember is that the total number of states, or energy levels, of a black hole can be estimated using simple arguments from thermodynamics, assuming that a black hole carries a temperature as given by Hawking[1]:

$$kT = 1/8\pi M\,, \tag{1.1}$$

in units where $c = h = G = 1$. The result is that the level density $\rho(M)$ as a function of the mass M is given by

$$\rho(M) = e^{4\pi M^2 + C}\,, \tag{1.2}$$

where C is an unknown constant. The point is that this number is small! If one counts the number of levels provided by the thermal particles in the vicinity of the black hole one finds that the particles further than about one Planck unit away from the horizon are sufficient to produce all the entropy corresponding to these levels. The ones closer to the horizon would provide an infinite contribution if we were allowed to use a linearized theory. Of course these particles do *not* obey a linearized theory, but the mechanism by which their contribution to the entropy is turned off is obscure.

For this reason we expect that incoming particles indeed do affect the details of the quantum state a black hole can be in, in the sense that they determine

states evolve into mixed states due to this uncertainty, but here this is clearly seen as a shortcoming in our information concerning the effective interactions. The uncertainty in question could be resolved for instance by performing accurate measurements.

details of those emerging particles that were closer to the horizon than one Planck length when the incoming particle entered. Our best guess is then that the black hole is just one set of the possible intermediate states in an S-matrix. It should be in no fundamental way different from ordinary particles. Light particles have a Schwarzschild radius much smaller than their Compton wavelength; for black holes this radius is much bigger. This distinction must be a gradual one. And so we arrive at the "S-matrix Ansatz"[5] for the black hole. Once we assume that the black hole has an S-matrix, we can actually derive many of its properties, because many of the relevant laws of physics are already known to us.

2. The Pseudostring

We first observe that the nature of the gravitational interactions between incoming and outcoming particles can very easily be characterized. Incoming particles produce a *horizon shift*. This horizon shift may be very tiny, but its effects upon the outgoing particles grow exponentially with time. They are also readily computable[6]. The wave functions of all outgoing particles are simply shifted, by an amount that depends on the angular location on the horizon.*

The quantum state is shifted, and hence the outgoing wave functions are all multiplied with factors $\exp(ip_{\text{out}}\delta y)$, where p_{out} is the momentum in Kruskal coordinates and δy the horizon shift, a function depending explicitly upon the angular coordinates ϑ and φ. The effect of this operation would be a harmless multiplication if the outcoming particles were in a Kruskal momentum eigenstate, but of course, in more relevant circumstances they are not in such eigenstates. This way we conclude that any alteration of the form

$$|\psi\rangle_{\text{in}} \to |\psi + \delta\psi\rangle_{\text{in}}\,, \tag{2.1}$$

where $\delta\psi$ carries a given momentum $p_{\text{in}}(\vartheta,\varphi)$, affects the outcoming state by the above given operation.

We can now repeat the argument as many times as we wish so that, in principle, we should obtain *all* other S-matrix elements. The procedure, and its results, are described in Ref[5]. They can be summarized as follows.

The momenta of in- and outgoing particles $p_{\text{in}}(\vartheta,\varphi)$ and $p_{\text{out}}(\vartheta,\varphi)$, are to be defined with respect to Kruskal coordinates, not Schwarzschild coordinates this is a point of concern, to be discussed later. When specified at all angular positions (ϑ,φ) these momenta, and in addition some other quantities such as electric charge density $\rho(\vartheta,\varphi)$, these variables should *entirely* specify the quantum states of the in- and out- quantum states respectively. So we refer to these states as

$$|p_{\text{in}}(\Omega)\,,\rho_{\text{in}}(\Omega)\rangle \quad \mathit{and} \quad |p_{\text{out}}(\Omega)\,,\rho_{\text{out}}(\Omega)\rangle\,, \tag{2.2}$$

* This angular dependence is crucial for our arguments, since without such an angular dependence one could transform (practically) all its effects away. This is why one must be very careful in interpreting some popular two-dimensional toy models of black holes[7].

where Ω stands for (ϑ,φ). The resulting S-matrix can then be written as a functional integral

$$\langle p_{\rm out}(\Omega)\,,\rho_{\rm out}(\Omega)\,|\,p_{\rm in}(\Omega)\,,\rho_{\rm in}(\Omega)\rangle =$$

$$\mathcal{N}\int\mathcal{D}u_\mu(\Omega)\mathcal{D}\phi(\Omega)\exp\, i\int d^2\Omega\Big(-\tfrac{1}{2}(\partial_\Omega u_\mu)^2+p_\mu u_\mu-\tfrac{1}{2\kappa}(\partial_\Omega\phi)^2+\phi(\rho_{\rm out}-\rho_{\rm in})\Big)\,. \qquad (2.3)$$

Here $\mathcal{N}$ is a normalization factor; the Lorentz index μ is defined such that $p_\mu = (\sqrt{2})^{-1}(p_{\rm in}+p_{\rm out}\,,0,0,p_{\rm in}-p_{\rm out})$, similarly u_μ, and κ is a constant defining the unit of electric charge. u_μ and ϕ are functional integration variables, depending on the two angular coordinates on the black hole horizon. u_μ are like the two transverse dynamic variables of a "string" whose world sheet is the intersection of the future and the past horizon of the black hole, which is a two-dimensional surface. ϕ is a periodic variable (it is defined as an angle *modulo* 2π). This is a consequence of electric charge quantization. Observe that in every respect electromagnetism appears to be represented here as a Kaluza-Klein theory. This was *not* put in but came out of our theory as a consequence of the S-matrix Ansatz.

The similarity between Eq. (2.3) and a string theory amplitude is striking. This resemblance becomes even closer if we represent the in- and out-state particles in wave-packets. One then has to integrate over the coordinates (ϑ,φ) convoluted with a wave function, and these integrals then correspond to the Koba-Nielsen integrations. An important *difference* between (2.3) and string theory is the factor i in the exponent, which corresponds to a purely imaginary string constant*. Our interpretation of this observation is that the black hole horizon can in some repects be regarded as the world sheet of a virtual closed string. The external particles are inserted there as vertex insertions in the usual sense.

We discovered that one can start with several kinds of fundamental interactions in one's favorite standard model and observe that these are reproduced in the functional integral (2.3) on the horizon. Electromagnetism, here represented by the variable ϕ, being just an example. Non-Abelian interactions give rise to more complex variables in two dimensions. Quite generally however the following picture emerges: The *gauge transformation generators* of the 4-dimensional theory correspond to the *dynamical variables* in the 2-dimensional one. Therefore the spin of a physical degrees of freedom in 2 dimensions is one less than the corresponding one in 4 dimensions.

Scalar and Dirac-spinor fields seem not to generate anything in 2 dimensions. An exception to this is the occurrence of spontaneous symmetry breaking: if in four dimensions a symmetry is broken spontaneously, the corresponding symmetry in 2 dimensions is *explicitly* broken: the scalar field in 4 dimensions maps into a "spurion" field in 2 dimensions (spurions were used in the '60's to describe explicit symmetry breaking interactions). Indeed one may view the value of the scalar fields at the horizon intersection point as being the spurion parameter.

* The fact that the string constant comes out imaginary should not be seen as a departure from unitarity, as was asserted by one author, but rather as a *consequence* of unitarity as required in our formalism.

A dual transformation in 4 dimensions corresponds to a similar dual transformation in 2 dimensions. Thus, magnetic monopoles entering the black holes generate a topological kink in the two-dimensional system; furthermore, quark confinement in 4 dimensions can be seen to correspond to an explicit symmetry breaking in terms of the scalar disorder parameter in two dim nsions.

Proceeding along these lines it is natural to suspect that a *gravitino* in four dimensions corresponds to a Dirac spinor in the 2-dimensional theory. What we have not understood at present however is how to incorporate effects of Dirac spinors in four dimensions in the 2-dimensional theory; they seem to leave no trace.

For more details of the string picture of black holes we refer to Ref[5].

3. Problems with Unitarity

Is our scattering matrix (2.3) unitary? A strange new problem arises. One may observe that the scattering matrix will indeed be unitary, but only so in a very unconventional Hilbert space. Two states that have exactly the same momentum (and charge) distribution for the ingoing – or outgoing – particles, cannot be distinguished any other way and therefore must be identical. In particular the *number* of particles entering or leaving at a given spot on the horizon cannot be specified. This implies that the Fock space of elementary particles will eventually look very different from what it used to be in elementary particle physics. For instance, the in- and out- states will carry no label specifying their baryon number. Consequently the black hole scattering matrix cannot possibly obey baryon number conservation. Clearly continuous global symmetries in our fundamental particle interactions cannot be reproduced in the black hole scattering matrix.

Another apparent problem with unitarity arises if the shift $\delta y(\vartheta,\varphi)$ at some values of ϑ and φ becomes too large. It could then be that a particle, originally destined to emerge in the out-state when the in- wave function was $|\psi\rangle$, is shifted beyond the horizon when the in-state is $|\psi+\delta\psi\rangle$. This is a consequence of the fact that we had been forced to define momenta in Kruskal coordinates in stead of Schwarzschild coordinates. A shift in Kruskal space can bring a particle behind the horizon.

We should stress that this latter problem is only an apparent one. There is no real contradiction with unitarity here because we imagine the total set of allowed out- states to be much smaller than the Hilbert space spanned by *all* possible waves of outgoing particles. The shift δy does not affect one single particle but an infinite series of particles emerging at all times. So if one or several of these disappear behind the horizon there are always enough others left to enable us to distinguish this shifted out- state from other out- states. Thus, our problem is more of a practical nature than fundamental. It tells us that the standard way to build up a Hilbert space in terms of plane wave of particles cannot be used here.

These problems must be related to another practical problem: even the set of all functions $p(\vartheta,\varphi)$ and $\rho(\vartheta,\varphi)$ is too large. Our entropy aruments suggest that there should be no more than about one Boolean variable per unit of surface area on the horizon in Planck units. This is as if these functions p and ρ have a cut-off. Components of their Fourier transforms in the transverse directions with momentum larger than a Planck unit should be removed or considered redundant. On the other hand lots of details on a distance scale just a bit larger than the Planck length are described by as yet unknown parts of the standard particles interactions. These details will be essential in the definitions of inner products in our Hilbert spaces, yet they are not yet accessible to us because the particle interactions at those scales are not yet known.

All this may seem to be extremely unconventional and inaccessible physics. But it is not quite that bad. We emphasize that the mathematical situation here is exactly as in string theories. In string theory also it is not the entire Hilbert space but rather the scattering matrix that is constructed. If particles are identified as vertex insertions on a string world sheet then exactly the same feratures do show up in string theory. Consider namely the Koba-Nielsen integrand with a given array of vertex insertions, for a given N particle amplitude. If in this integrand two vertex insertions occur at the same spot on the world sheet then this is indistinguishable from the integrand for the $N-1$ particle amplitude. Replace the string world sheet by the horizon. The indistinguishability of two particles on the same spot on the horizon, or rather the fact that this state cannot be distinguished from a single particle state at that spot, has the same mathematical origin.

4. Unitarity in Complete Black Hole Histories

Our scattering matrix Ansatz tells us to assume as a starting point the existence of a scattering matrix for a black hole. And then we can deduce information about this matrix by applying all physical laws we know. The only reason why this does not work completely is that we only know the interactions between elementary particles at low energies, or, equivalently, at large distance scales. So we do not know how to characterize the very small distance features of our scattering matrix, and since inner products of states depend crucially also on the small distance features, we run into problems as described in the previous section. The general strategy we are trying to implement is to use the known laws in as many forms as possible to reduce these uncertainties as much as possible. Also we can try certain assumptions concerning the small distance interactions to check which of these produce a consistent theory (we saw for instance that baryon number conservation must not be a symmetry of our basic interactions).

With this strategy in mind we now proceed to consider a branch of the scattering matrix different from the one considered before, namely the transition amplitude from a black hole just formed into a black hole exploding into expanding dust shells. Thus we consider a completely specified in-state, $|\mathrm{in}\rangle$, a completely specified out-state, $|\mathrm{out}\rangle$, and assume that one single amplitude $\langle\mathrm{out}|\mathrm{in}\rangle$ is given. As before, the question is to deduce other amplitudes

$$\langle\mathrm{out}+\delta_{out}|\mathrm{in}+\delta_{in}\rangle\,, \tag{4.1}$$

where δ_{out} and δ_{in} are tiny alterations. We now proceed in a way very different from Section 2, namely by first postulating a singularity free, topologically trivial space-time metric corresponding to the original amplitude. That this is possible at all is surprising and requires some discussion. The trick is to assume the outcoming matter to be due to some unspecified interaction process very near the horizon which gave rise to extremely strong curvature there. This curvature would not be directly detectable for ingoing or outgoing observers and therefore its presence does not contradict anything we know. If it *were* observable it *would* contradict the ordinary laws of physics. This is a necessary aspect of the S-matrix Ansatz. The mere assumption that an amplitude $\langle \text{out}|\text{in}\rangle$ exists does contradict "normal" laws of physics. So we are forced to assume something out of the ordinary there, and the least harmful way to do this is to postulate a *conical singularity* (actually it is not a singularity but just a region of very strong curvature, because the singularity will be slightly smoothened). The presence of such a singularity will only be visible when devices sent into the (approximate) black hole and reappearing somewhat damaged after the black hole decayed, are compared to apparatus that stayed just outside. But such an experiment will be impossible classically. In stead of these "devices" we will just consider infinitesimal additions $\delta\psi$ to our wave functions and study the effects on these.

The Penrose diagram is now the one pictured in Fig. 1a. It is topologically trivial. Apart from a mild (very slightly smeared) singularity at the point S there are no further singularities. The dotted lines are very much like horizons, but of course they are not horizons, they replace them. At the point S the standard laws of physics seem to be not obeyed. The curvature there is the one produced by a very violent "interaction" that caused the incoming shell of matter to turn around and go outwards. It is as if a "chemical" explosion takes place there which was just strong enough to avert the gravitational implosion. Let us stress again that an observer who stays outside the black hole (or "pseudo-black-hole") can never detect this curvature, so that from his point of view all laws of physics are obeyed.

What we claim now is that this proces may well be reconciled completely with the known loaws of physics, even at S, by studying quantum field theoretical effects caused by the curvature at S. In the next Section we shall prove that the singularity is such that if one starts off with a local vacuum, a nearly infinite spectrum of particles will be created there. We will then argue that if on the dotted lines in Fig. 1. we require the *absence* of particles, there *must* be particles in the gray area. Originally we had "postulated" that there are particles there; we can now derive that the postulate may well be correct. So the whole picture may become self-consistent.

In our simplified model we replace all incoming and all outcoming matter by single "dust shells". Upon careful inspection one finds that this is hardly an approximation, see Fig. 1, where all matter coming in is squeezed towards the "far past" and everything coming out towards the "far future". Near S the most regular coordinate frame is a "temporary" Kruskal frame, and hence all matter in our space-time diagram is very strongly Lorentz boosted.

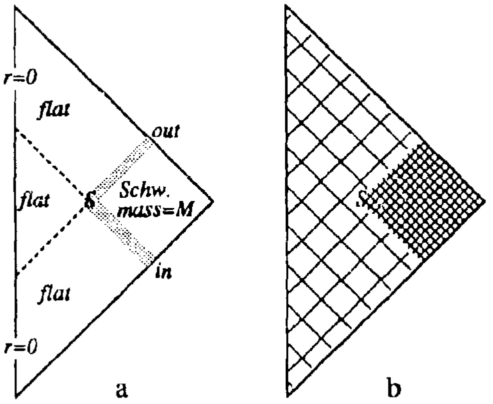


Fig. 1. a) Non-singular Penrose diagram for an entire black hole history.
b) Coordinate frame in the outside region of the black hole is more dense. At the point S there is a conical singularity.

Our model then is of the same type as the some of the systems studied by T. Dray and the author in Ref[8]. Here we studied the effect of in- and outcoming dust shells on the Schwarzschild metric. But in this work we explicitly postulated the absense of conical singularities at points such as S, so that the occurance of typical black hole singularities at $r = 0$ both in the past and in the future is inevitable. Now we take the same models *with* conical singularity at S, chosen in such a way that the singularities at $r = 0$ go away. The metric one then gets fits naturally with the S-matrix Ansatz. The strategy is now simple. In Fig. 1 we postulate space-time to be flat in all of the interior region, except in the quadrant where an outside observer sees the black hole. There we have the Schwarzschild metric corresponding to a mass M. Consider the Kruskal coordinates x and y . Let the physical quadrant be given by $x > 0, y > 0$. Very near the Schwarzschild horizons, at the line $x = x_0$ and the line $y = y_0$, where x_0 and y_0 are very small but positive, we have the matter shells. At those shells we glue the Schwarzschild metric against the flat space-time metric such that the Schwarzschild r parameter matches with the flat space r parameter. The metric is then C^0.

But a singularity develops at S . This we see as follows. Suppose we use Penrose coordinates, that is, coordinates such that the local light cones have a width of exactly 45°. One then finds that the gluing procedure just described forces us to scale down the Schwarzschild solution (as written in Kruskal coordinates) to a very small size, and to blow up the internal region of the black hole to large sizes. This is sketched in Fig. 1b by drawing dense coordinate lines in the outside region and wide coordinate lines inside.

In Fig. 2 we illustrate what happens to geodesics near such a point. At the point B in Fig. 2c we make the transition to Lorentz transformed coordinates,

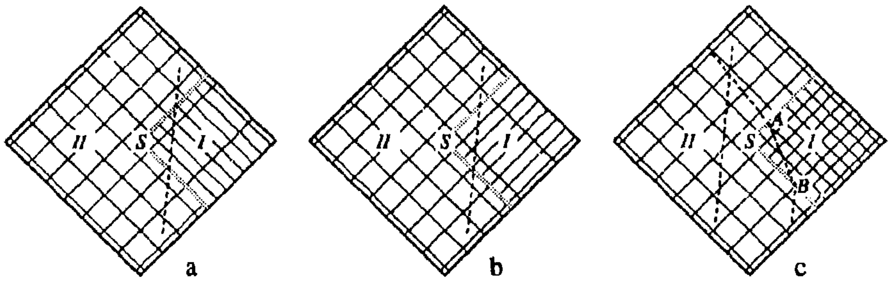


Fig.2. In a) and b) the region I is Lorentz boosted compared to region II. Since this is just a coordinate reparametrization a geodesic (dotted line) goes straight. In c) we performed the transformation of a) at the point A and the one of b) at the point B. At both points, the orthogonal coordinate was squeezed. Consequently, a geodesic going through region I is now bent over.

but because the orthogonal coordinate is scaled in the wrong direction a geodesic crossing at B is bent over. The same happens in A. Thus, two particles with equal velocities may end up having different velocities if they pass the point S at opposite sides. Thus the singularity at S has the effect of a Lorentz transformation if one follows a loop around it.

For a black hole with lifetime long compared to its size the lorentz boost across S is extremely large. For the remainder of our considerations we prefer to concentrate on the case that this Lorentz boost is not so extremely large. This happens either if one considers very tiny black holes, or black holes with an extremely "unlikely" history. The only reason then why this history is unlikely for large black holes is that the amplitude is too small after multiplication with the appropriate phase space factor, which is also too small, so that other processes (giving the hole a lifetime of order M^3) are more probable. We just point out that this is not at all an objection against considering the amplitudes for such "unlikely" histories.

Thus, we concentrate on Fig. 1 where the region very close to the origin, S, is described by Fig. 2c. let the total Lorentz boost along a closed curve be given by the parameter ϕ in the boost matrix

$$L = \begin{pmatrix} \cosh\phi & \sinh\phi & 0 & 0 \\ \sinh\phi & \cosh\phi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \tag{4.2}$$

The local effect of the shells of matter is small compared to the effect of the conical singularity.

5. Particle Creation by a Conical Singularity

We now consider the effect a conical singularity of the sort described in the previous Section has on a quantized state in field theory. Since the metric has no timelike Killing vector there is no conserved energy. If we begin with the vacuum state at $t = -\infty$ the state at $t = +\infty$ will in general contain particles. The computation is not hard. Observe that, in contrast with the familiar calculation of the Hawkirg-Unruh effect there will be no information loss. Later we will be interested in different initial states, but let us begin with the vacuum.

For simplicity we take the field to be scalar. The local operators $\varphi(\mathbf{x},t)$ and $\dot\varphi(\mathbf{x},t)$ are given by

$$\varphi(\mathbf{x},t) = \sum_k \frac{1}{\sqrt{(2Vk^0)}}(a_k e^{ikx} + a_k^\dagger e^{-ikx}) , \tag{5.1}$$

and

$$\dot\varphi(\mathbf{x},t) = \sum_k \sqrt{\frac{k^0}{2V}}(-ia_k e^{ikx} + ia_k^\dagger e^{-ikx}) , \tag{5.2}$$

where a_k and $a_k^\dagger$ are annihilation and creation operators at given three-momentum $\mathbf{k}$. As usual we define $k^0 = \sqrt{(\mathbf{k}^2 + m^2)}$ and $kx = \mathbf{kx} - k^0 t$.

We take Eqs (5.1) and (5.2) to hold at time $t < 0$, before the singularity S occurred. At time $t > 0$ we take the fields to be

$$\varphi(y) = \sum_k \frac{1}{\sqrt{(2Vk^0)}}(b_k e^{iky} + b_k^\dagger e^{-iky}) , \tag{5.3}$$

where y are Cartesian space-time coordinates at $t > 0$. They are related to the $\mathbf{x}, t$-coordinates by

$$y = x \quad \text{if} \quad x_1 < 0 , \qquad y = L^{-1}x \quad \text{if} \quad x_1 > 0 , \tag{5.4}$$

where L is the Lorentz transformation (4.1).

One finds that

$$b_p = \sum_k A^+_{pk} a_k + \sum_k A^-_{pk} a_k^\dagger , \tag{5.5}$$

where A^+_{pk} and A^-_{pk} are coefficients. From now on the variables p and k are only the x-components of the momenta, the ones that transform non-trivially under the Lorentz transformation (4.1). p^0 and k^0 are the usual time components of the momenta. Also we write $x = x_1$. Let us furthermore use the shorthand notation

$$\cosh\phi = c \quad , \quad \sinh\phi = s , \tag{5.6}$$

where ϕ is the Lorentz boost parameter. We will use a finite-volume formulation so that the momenta are discrete. The coefficients are then computed to be

$$A^\pm_{pk} = \frac{1}{2V}\sqrt{\frac{p^0}{k^0}} \int_0^\infty dx \left((1 \pm \frac{k^0}{p^0})\, e^{-i(k-p)x} + (1 \pm \frac{ck^0 - sk}{p^0})\, e^{i(ck - sk^0 - p)x} \right) , \tag{5.7}$$

where V is the volume (soon to be sent to infinity).

The integral over x can of course be calculated:

$$A^\pm_{pk} = \frac{1}{2V\sqrt{p^0k^0}} \left(\frac{-i(p^0 \pm k^0)}{k - p - i\epsilon} + \frac{i(p^0 \pm (ck^0 - sk))}{ck - sk^0 - p + i\epsilon} \right) . \tag{5.8}$$

It is illustrative to compute the occupation number $\langle b_p^\dagger b_p\rangle_0$, where $\langle\rangle_0$ corresponds to the vacuum of the annihilation operators a_k. It is found to be

$$\begin{aligned}\langle b_p^\dagger b_p\rangle_0 &= \sum_k |A_{pk}^-|^2 = \\ &= \frac{1}{8\pi V p^0}\int\frac{dk}{k^0}\left(\frac{(p^0-k^0)(ck-sk^0-p)+(p-k)(p^0-ck^0+sk)}{(k-p)(ck-sk^0-p)}\right)^2 ,\end{aligned} \tag{5.9}$$

where the summation was replaced by the integral for $V \to \infty$, and in the integral we must insert

$$k^0 = \sqrt{k^2+\tilde{k}^2+m^2}\,, \tag{5.10}$$

and similarly for p^0. Here $\tilde{k}$ is the transverse part of the momentum k.

The rest is straightforward arithmetic. All integrals can be performed and the result is

$$\langle b_p^\dagger b_p\rangle_0 = \frac{1}{2\pi V p^0}\left[\frac{\phi}{\tanh\frac{1}{2}\phi}-2\right]. \tag{5.11}$$

For small ϕ the quantity between square brackets is

$$[\ldots] = \frac{\phi^2}{6}-\frac{\phi^4}{360}+\ldots. \tag{5.12}$$

and if ϕ is large then it approaches

$$[\ldots] = |\phi|-2+2|\phi|e^{-|\phi|}+\ldots \tag{5.13}$$

Note that the p dependence is $d^3p/2p^0 = d^4p\delta(p^2+m^2)$, which is Lorentz invariant. Invariance under Lorentz transformations in the x direction is not surprising. But the invariance in the transverse direction is an accident. The coefficients $A^\pm$ themselves do not have this latter invariance. Also, the fact that Eq. (5.11) is independent of the sign of ϕ is an accident.

The coefficients of Eq. (5.8) were computed for given 3-momenta. The calculations simplify however if we go to lightcone coordinates instead. The outcome, such as Eq. (5.11), of course stays the same.

6. Conclusion

We propose to use the metric of Fig. 1 to compute amplitudes (4.1) if one single amplitude $\langle\text{out}|\text{in}\rangle$ is given. The conical singularity S is not strong enough to cause any loss of information. If S were infinitely sharp a vacuum in-state would cause an unlimited particle production into the out-state. We can put a bound on this particle production by smearing the singularity a bit. We showed that calculating the evolution of the state that started out as a local vacuum is straightforward. But what actually will be needed is the evolution of a state that has no particles coming from $(r = 0)$ (the lower dotted line in Fig. 1a) into a state that has no particles moving towards $(r = 0)$ (the upper dotted line in Fig. 1a). In general these states may have particles on the other side (the gray areas in Fig. 1a). Computation of these transitions is much harder because the distinction between left-goers and right-goers is to some extent arbitrary and hence difficult to implement.

We note that the in- and outcoming particles caused by the singularity essentially imply that our Ansatz for the metric is self-consistent. Somewhat more precisely, we propose the following. In contradistinction to the procedure we proposed previously, and which was recapitulated in Section 2, we now assume that the variations $\delta\psi$ of both incoming and outgoing states are too small to have any direct gravitaional effect, so that here we can superimpose quantum states in the usual way. We will refer to the particles in $\delta\psi$ as "soft" particles. All particles whose gravitational effects we wish not to ignore (the "hard" particles) we put in the original states $|\text{out}\rangle$ and $|\text{in}\rangle$. So in each of these "gravitational windows" we can compute a block

$$\langle \text{out} + \delta_{out} | \text{in} + \delta_{in} \rangle \,. \tag{6.1}$$

Indeed all these amplitudes are uniquely defined up to one overall multiplicative constant. There is no drain of information. On the other hand however, there is a divergence: if S is infinitely sharp the majority of transitions will contain huge numbers of particles modifying the already heavily populated in- and out-states. Just because we wish to consider only soft particles in $\delta\psi$ we must accept a cut-off for the singularity S. The exact location of the cut-off, the transition region between soft and hard particles, must to some extent be irrelevant.

Note that the transitions $\delta_{in} \to \delta_{out}$ themselves will not violate any of the symmetries of our standard interactions. However in the entire block (6.1) the hard particles will violate all global symmetries, but for the entire block this violation will be the same.

We believe [illegible] our new proposal will open up different elements of the black hole scattering matrix and allow us to study this matrix further. Ultimately all procedures should be combined into one single theory, but we are not yet that far. By construction it seems that there cannot be any violation of unitarity for this matrix, but we should admit that this has not yet been demonstrated. The problem is now that the S-matrix describing the soft particles alone, after the cut off, will be unitary. But without cut-off the blocks (6.1) that we have are each *different* parts of *diferent* S-matrices. Each of theses matrices separately are unitary, but whether this combination will again be unitary remains to be seen. A delicate study of the various limiting procedures involved will be needed to answer such questions.

Acknowledgement

The author benefited from numerous discussions with his colleagues, in particular B. Whiting, L. Susskind and Chr. Stephens.

References

1. S.W. Hawking, *Commun. Math Phys.* **43** (1975) 199; J.B. Hartle and

S.W. Hawking, *Phys. Rev.* **D13** (1976) 2188;
S.W. Hawking and R. Laflamme, *Phys. Lett.* **B209** (1988) 39.

2. S.W. Hawking,*Phys. Rev.* **D14** (1976) 2460; *Commun. Math. Phys.* **87** (1982) 395; D.N. Page, *Phys. Rev. Lett.* **44** (1980) 301;
   D.J. Gross, *Nucl. Phys.* **B236** (1984) 349.
3. S.W. Hawking and G.F.R. Ellis, *"The Large Scale Structure of Space-Time* (Cambridge Univ. Press, 1973);
   J.D. Bekenstein, *Nuovo Cim. Lett.* **4** (1972) 737; *Phys. Rev.* **D7** (1973) 2333; **D9** (1974) 3292
4. G. 't Hooft, *Nucl.Phys.* **B256** (1985) 727
5. G. 't Hooft, *Nucl Phys.* **B335** (1990) 138; *Physica Scripta* **T15** (1987) 143, *ibid.* **T36** (1991) 247; G. 't Hooft *in "Black Hole Physics"*, NATO ASI Series C, Vol. 364, p. 381, De Sabbata and Zhang eds, Kluwer; *and in "New Symmetry Principles in Quantum Field Theory"*, NATO ASI Series B, Vol. 295, p. 275, J. Frölich et al eds, Plenum.
6. T. Dray and G. 't Hooft, *Nucl. Phys.* **B253** (1985) 173.
7. V.P. Frolov et al, *Phys. Lett.* **B 279** (1992) 29; C.G. Callan et al, *Phys. Rev.* **D 45** (1992) R1005; A. Peet et al, Stanford University Preprint SU-ITP-92 16 (May 1992).
8. T. Dray and G. 't Hooft, *Commun. Math. Phys.* **99** (1985) 613

ENTROPY GENERATION IN QUANTUM GRAVITY AND BLACK HOLE REMNANTS[*]

To our friend Yakir for his birthday

A. CASHER and F. ENGLERT[†]

Service de Physique Théorique
Université Libre de Bruxelles, Campus Plaine, C.P.225
Boulevard du Triomphe, B-1050 Bruxelles, Belgium
and
School of Physics and Astronomy
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel

ABSTRACT

The area entropy $A/4$ and the related Hawking temperature in the presence of event horizons are rederived, for de Sitter and black hole topologies, as a consequence of a tunneling of the wave functional associated to the classical coupled matter and gravitational fields. The extension of the wave functional outside the barrier provides a reservoir of quantum states which allows for an additive constant to $A/4$. While, in a semi-classical analysis, this gives no new information in the de Sitter case, it yields an infinite constant in the black hole case. Evaporating black holes would then leave residual "planckons" - Planckian remnants with infinite degeneracy. Generic planckons can neither decay into, nor be directly formed from, ordinary matter in a finite time. Such opening at the Planck scale of an infinite Hilbert space is expected to provide the ultraviolet cutoff required to render the theory finite in the sector of large scale physics.

1. Introduction

Tunneling in quantum gravity can generate entropy[1,2]. To understand how such an apparent violation of unitarity may arise, let us first consider a classical spacetime background geometry with compact Cauchy hypersurfaces. If quantum fluctuations of the background are taken into account, quantum gravity leaves no "external" time parameter to describe the evolution of matter configurations in this background. Indeed, the solutions of the Wheeler-De Witt equation[3]

$$\mathcal{H}|\Psi\rangle = 0 \tag{1}$$

where $\mathcal{H}$ is the Hamiltonian density of the interacting gravity matter system can contain no reference to such time when there are no contribution to the energy from

[*] presented by F.Englert

[†] E mail: casher at taunivm; fenglert at ulb.ac.be

surface terms at spatial infinity[4]. This is due to the vanishing of the time displacement generator and even though the theory can be unambiguously formulated only at the semi-classical level, such a consequence of reparametrization invariance should have a more general range of validity.

To parametrize evolution, one then needs a "clock" which would define time through correlations[5], namely correlations of matter configurations with ordered sequences of spatial geometries. If quantum fluctuations of the metric field can be neglected, the field components g_{ij} at every point of space can always be parametrized by a classical time parameter, in accordance with the classical equations of motion. This classical time, which is in fact a function of the g_{ij}, can be used to describe the evolution of matter and constitutes thus such a dynamical "clock" correlating matter to the gravitational field[6]. This description is available in the Hamilton-Jacoby limit of Eq.(1) where the classical background evolving in time is represented by a coherent superposition of "forward" waves formed from eigenstates of Eq.(1). When quantum metric field fluctuations are taken into account, "backward" waves, which can be interpreted as flowing backwards in time, are unavoidably generated from Eq.(1) and the operational significance of the metric clock gets lost outside the domain of validity of the Hamilton-Jacoby limit. Nevertheless, in domains of metric field configurations where both forward and backward waves are present but where quantum fluctuations are sufficiently small, interferences with such "time reversed" semi-classical solutions will in general be negligible[‡]. Projecting then out the backward waves restores the operational significance of the metric clock but the evolution marked by the correlation time is no more unitary: information has been lost in projecting these backward waves stemming from regions where quantum fluctuations of the clock are significant. This is only an apparent violation of unitarity which would be disposed of if the full content of the theory would be kept, perhaps eventually by reinterpreting backward waves in terms of the creation of "universe" quanta through a further quantization of the wavefunction Eq.(1).

This apparent violation of unitarity is particularly marked if the gravitational clock experiences the strong quantum fluctuations arising from a tunneling process. This can be illustrated from the simple analogy, represented in Fig.1, offered by a nonrelativistic closed system of total fixed energy where a particle in one space dimension x moving in a potential $U(x)$ plays the role of a clock for surrounding matter and tunnels through a large potential barrier. Outside the barrier, the clock is well approximated by semi-classical waves, but if on the left of the turning point a one would take only forward waves, one would inevitably have on the right of the other turning point b both forward and backward waves with large amplitudes compared with the original ones. The ratio between the squares of the forward amplitudes on the right and on the left of the barrier for a component of the clock wave with given clock energy E_c is the inverse transmission coefficient $N_0(E_c)$ through the barrier and provides a measure of the apparent violation of unitarity. In the Hamilton-Jacoby limit of quantum gravity, the characterisation of tunneling amplitudes by inverse transmission coefficients N_0 will appear as the natural one to compute the entropy transferable reversibly between the metric clock and matter. More precisely, we shall

[‡] For a recent discussion of related problems see reference 7.

see that, in this limit, spherically symmetric spacetimes bounded by event horizons are in general connected by tunneling to another manifold and that the entropy gained by tunneling from the latter to the former is, for large barriers, $\log N_0$. Explicit evaluation of this tunneling entropy yields $\ln N_0 = A/4$ where A is the area of the event horizon. In this way the horizon thermodynamics of Gibbons and Hawking[8] is recovered. But the present approach has potentially additional information.

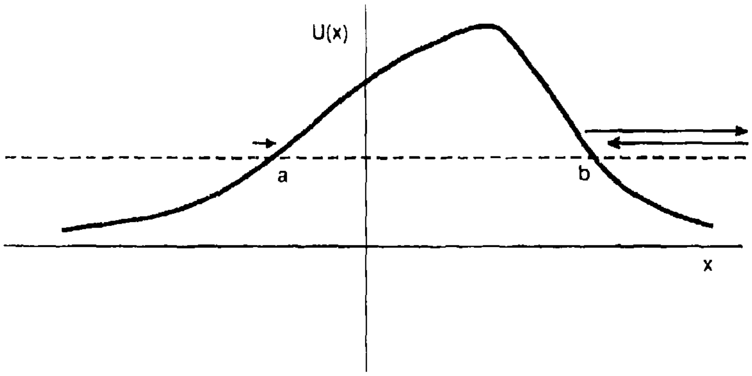


Figure 1. Tunneling of a nonrelativistic "clock". The energy of the clock E_c is represented by the dashed line. On the left of the turning point a the clock is well represented by a forward wave depicted here by a single arrow. On the right of the turning point b the amplification of the forward wave and the large concomitant backward wave are indicated.

The tunneling entropy $\ln N_0$ is in last analysis an effect of quantum fluctuations in quantum gravity. Therefore, despite the fact that no violation of unitarity would appear in a complete description including backward waves, this entropy should be expressible in terms of density of states of matter and gravity. Tunneling offers an interesting perspective in this direction because it enlarges the semi-classical wave function of spacetime to include in its description the other side of the barrier. This can yield a reservoir of quantum states which may provide, in addition to the $\exp(A/4)$ states building the entropy, residual states which would be expressed as an "integration constant" in the total entropy S of spacetime. Thus we shall write

$$S = A/4 + C \tag{2}$$

and try to get some information about the constant C by analysing both sides of the barrier.

The knowledge of C is crucial, in particular for the understanding of the black hole behaviour at the final stage of evaporation.

A infinitely large value of C would indeed indicate that the evaporation can only radiate a finite number of "surface" states of order $\exp(A/4)$ out of an infinite set of available internal states. This mismatch would entirely modify the black hole evaporation process at its last stage and bring the decay to a halt. Indeed when the black hole evaporates to the Planck scale, it becomes, if C is infinite, a "planckon"[9], that is a remnant with infinite degeneracy. Causality and unitarity prevent the decay and the production in a finite time of planckons directly out of ordinary matter for nearly all such states[9] Namely, if a planckon state $|A_i\rangle$ of finite size and mass m decays (or is produced) within a finite time τ in an approximately flat space time background, the total number of possible final (or initial) states is limited through causality by the number $\mathcal{N}(\tau)$ of orthogonal states with total mass m in a volume τ^3. Assuming the number of quantum fields which describe physics at scales large compared to the Planck scale to be finite, $\mathcal{N}(\tau)$ is a finite number. Unitarity then implies that if the dimension $\nu(m)$ of the Hilbert space spanned by the degenerate states $|A_i\rangle$ becomes greater than $\mathcal{N}(\tau)$, a subspace of planckon states whose dimension is $\nu(m) - \mathcal{N}(\tau)$ will be, for times smaller than τ, orthogonal to the Hilbert space of states formed by these quantum fields. Thus, when $\nu(m) \to \infty$, generic planckons cannot decay (nor be formed) in a finite time. Of course the above argument does not preclude the very formation of the *finite* number of distinct planckons which can be generated in a finite time as remnants of macroscopic decaying black holes. This time is however unrelated to the time for their decay (creation) directly at the Planck size into (from) ordinary matter quanta; the latter time is generically infinite.

On the other hand, a zero or finite value of C would lead to the disappearance of the hole at the end point of evaporation and hence probably imply a genuine violation of unitarity within our universe*.

The main content of our work is that, in an asymptotically flat background, the constant C for black holes, as deduced from a WKB analysis of tunneling, is in fact infinite. Hence the present approach indicates that the solution of the unitarity problem posed by the black hole decay is provided by planckon remnants. This conclusion is however contingent upon the limitation of the semi classical approach to quantum gravity used here and remains therefore a tentative one.

We shall first review the computation of tunneling amplitudes in quantum gravity through static barriers[2] and compute the tunneling entropy for the case of de Sitter spacetime topologies. However the estimation of C appears in this case intractable within the semi classical approximation. We then shall analyse in similar terms the black hole geometries[12]. The above mentioned results will be derived and discussed.

* For a comprehensive review on recent attempts to solve the black hole unitarity puzzle, see reference 10. See also reference 11

2. Tunneling amplitudes in quantum gravity.

Our basic action in four dimensional Minkowski space-time will be

$$S = S_{grav} + S_{matter} \tag{3}$$

where S_{grav} has the conventional form ($G = 1$) :

$$S_{grav} = -\frac{1}{16\pi} \int \sqrt{-g} R \, d^4x \tag{4}$$

and S_{matter} contain sufficiently many free parameters to allow for the stress tensors considered below. A possible cosmological constant term can be included in the matter action.

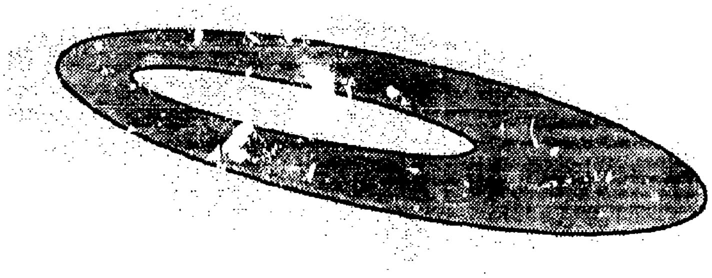


Figure 2. Tunneling in quantum gravity. The two solid curves represent turning hypersurfaces Σ_1 and Σ_2 separating the dark gray Euclidean region $\mathcal{E}$ from two Minkowskian spacetimes depicted in light gray.

Consider (Fig.2) in general two spacelike hypersurfaces Σ_1 and Σ_2 which are turning points in superspace (or turning hypersurfaces) along which solutions of the Minkowskian classical equations of motion for gravity and matter meet a classical solution of their Euclidean extension. Σ_1 and Σ_2 are thus the boundaries of a region $\mathcal{E}$ of Euclidean space-time defined by the Euclidean solution. If $\mathcal{E}$ can be continuously shrunk to zero one can span $\mathcal{E}$ by a continuous set of hypersurfaces τ_e = constant such that $\tau_e \equiv \tau_{e,1}$ on Σ_1 and $\tau_e \equiv \tau_{e,2}$ on Σ_2. These τ_e = constant surfaces define a Euclidean coordinate system which we shall call synchronous; the Euclidean metric in $\mathcal{E}$ can be written in the form

$$ds^2 = N^2(\tau_e, x_k)\, d\tau_e^2 + g_{ij}(\tau, x_k)\, dx^i\, dx^j \tag{5}$$

where $N(\tau_e, x_k)$ is a lapse function.

The Euclidean action S_e over $\mathcal{E}$, from Σ_1 to Σ_2, is obtained by analytic continuation from the Minkowskian action Eq. (3) and can be written as

$$S_e(\Sigma_2,\Sigma_1) = \int\limits_{\mathcal{E}} \Pi^{ij} g'_{ij}\, d^4x + \int\limits_{\mathcal{E}} \Pi^a \Phi'_a\, d^4x - \int\limits_{\mathcal{E}} (g_{ij}\Pi^{ij})'\, d^4x \\ - \frac{1}{8\pi}\int\limits_{\mathcal{E}} \partial_k[(\partial_j N) g^{kj}\sqrt{g^{(3)}}]\, d^4x. \tag{6}$$

Here Π^{ij} and Π^a are the Euclidean momenta conjugate to the gravitational fields g_{ij} and to the matter fields ϕ_a; $g^{(3)}$ is the three dimensional determinant and the $'$ symbol indicates a derivative with respect to τ_e. In the gauges of Eq.(5), Π^{ij} is expressed as

$$\Pi^{ij} = \frac{\sqrt{g^{(3)}}}{32\pi N}[g^{im}g^{jn} - g^{ij}g^{mn}]g'_{mn}, \tag{7}$$

On the turning hypersurfaces Σ_1 and Σ_2, all field momenta (Π^{ij}, Π^a) are zero in the synchronous system and the third term in Eq.(6) vanishes. The last term in Eq.(6) also vanishes if the hypersurfaces Σ_1 and Σ_2 are compact but may receive contributions from infinity otherwise. In this case, we shall assume that turning hypersurfaces merge at infinity sufficiently fast so that the Euclidean action $S_e(\Sigma_2,\Sigma_1)$ does not get contributions from the last term in Eq.(6). The classical Minkowskian solution in the space-time $\mathcal{M}_1$ containing Σ_1 can be represented quantum mechanically by a "forward wave" solution $\Psi(g_{ij}, \phi_a)$ of the Wheeler-De Witt Eq.(1) in the Hamilton-Jacoby limit. At Σ_1, this wave function enters, in the WKB limit, the Euclidean region $\mathcal{E}$ and leaves it at Σ_2 to penetrate a new Minkowskian space-time $\mathcal{M}_2$. The tunneling of $\Psi(g_{ij}, \phi_a)$ through $\mathcal{E}$ engenders in addition to the "forward wave" solution a time reversed "backward wave". The inverse transmission coefficient N_0 through the barrier measures the ratio of the norms of the forward waves at Σ_2 and Σ_1. For large N_0 one may write in the synchronous system

$$N_0 = \exp\left[-2(\int\limits_{\mathcal{E}} \Pi^{ij} g'_{ij}\, d^4x + \int\limits_{\mathcal{E}} \Pi^a \phi'_a\, d^4x)\right]. \tag{8}$$

As all surface terms in Eq.(6) vanish in this system, Eq.(8) can be rewritten in the coordinate invariant form

$$N_0 = \exp[2S_e(\Sigma_1, \Sigma_2)]. \tag{9}$$

Let us examine the case where the Euclidean manifold $\mathcal{E}$ is static in the sense that it admits a Killing symmetry. We can take advantage of the covariance of the action S_e and express it in terms of a new "static" coordinate system, possibly singular, with momenta everywhere vanishing in $\mathcal{E}$. In this way, momenta in Eq.(8) get squeezed into the last surface term of Eq.(6) and one gets

$$N_0 = \Delta t_e \frac{1}{4\pi} \int_{\mathcal{E}} \partial_k[(\partial_j N) g^{kj} \sqrt{g^{(3)}}]\, d^3x \tag{10}$$

where Δt_e is the Euclidean time needed to span $\mathcal{E}$ in the static system. The tunneling amplitude will then be computable from this surface term only, even when the static parametrisation is singular.

3. The tunn ling entropy in de Sitter spacetime topologies

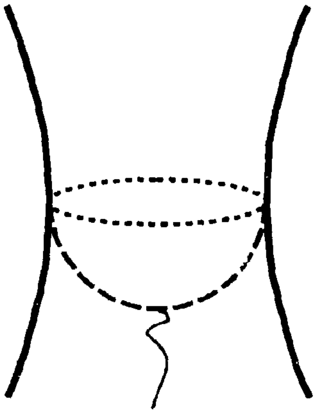


Figure 3. Tunneling in de Sitter topology. The heavy solid line delineates a 4-hyperboloid and the thin one a wormhole. The Euclidean domain $\mathcal{E}$ constituted by a half 4 sphere is delineated by a dashed line. The dotted circle is the turning hypersurface $\tau = 0$

Let us first illustrate the equivalence implied by Eq.(9) of the Eq.(8) and Eq.(10) for the de Sitter spacetime which is the classical solution of pure gravity in the presence of a cosmological constant Λ. In the present formalism, Λ should be viewed as the Lagrangian density of the matter action in Eq. (2); it plays the role of a matter distributed with rest energy density $\sigma = \Lambda$ and obeying the equation of state $\sigma = -p$ where p is a (negative) pressure. The full Minkowskian solution is the 4 hyperboloid (Fig.3) which can be parametrized by the minisuperspace metric

$$ds^2 = -d\tau^2 + a^2 d\sigma^2; \quad a = r_h \cosh\frac{\tau}{r_h} \tag{11}$$

where $r_h = (3/8\pi\Lambda)$. The hypersurface $\tau = 0$ is a turning hypersurface connecting the hyperboloid to the Euclidean solution consisting of the 4-sphere which can be described in a synchronous system by replacing in Eq.(11) τ by $-i\tau_e$. This yields the Euclidean scale factor

$$a_e = r_h \cos \frac{\tau_e}{r_h}. \tag{12}$$

The half-sphere delimited by $-\pi r_h/2 \leq \tau_e \leq 0$ has another turning point at, say, the south pole $\tau_e = -\pi r_h/2$ where $a_e = 0^*$: in the above synchronous system the space integral of the momenta in a τ_e = constant hypersurface vanishes in the vicinity of this point and so does the third term in Eq.(6). The half-sphere considered constitute the domain $\mathcal{E}$ through which a "wormhole" at, say, $a_e = 0$ is connected by tunneling to the de Sitter spacetime. The inverse transmission coefficient N_0 can be straightforwardly computed from Eq.(8) using Eq.(7) and one gets

$$N_0 = \exp\left[\frac{3\pi}{2} \int_{-\pi r_h/2}^{+\pi r_h/2} a_e \left(\frac{da_e}{d\tau_e}\right)^2 d\tau_e\right] = \exp(\pi r_h^2) = \exp(A/4) \tag{13}$$

where A is the area of the event horizon.

The significance of this result is best appreciated when the 4-sphere is described in static coordinates:

$$ds^2 = (1 - r^2/r_h^2)\,dt_e^2 + (1 - r^2/r_h^2)^{-1}\,dr^2 + r^2\,d\Omega^2. \tag{14}$$

In this static frame, all momenta vanish everywhere on the sphere and the tunneling is expressible by the surface term Eq.(10) only where the radial integration is carried from $r = 0$ to $r = r_h$. The Euclidean time is periodic with period $\mathcal{T}^{-1} = 2\pi r_h$. Using Eq.(10) with $\Delta t_e = (1/2)\mathcal{T}^{-1} = \pi r_h$, one recovers the result Eq.(13).

It is now easy to verify that the equality between the inverse transmission coefficient and $\exp(A/4)$ is maintained when the de Sitter spacetime is perturbed by spherically symmetric static matter distributions[2]. This establishes the validity of Eq.(13) for these generalized de Sitter spacetimes.

Let us tentatively take boundary conditions in field space by assigning pure forward waves at the wormhole turning point. The probability of finding an expanding generalized de Sitter spacetime for a corresponding wormhole state is then N_0, since in the classical limit interferences between spaces evolving forward or backward in time must be negligible. Assuming that all wormhole states are equally probable, we get from Eq.(13) that the relative probability of finding two matter configurations

[*] In fact, any point on the half-sphere can be taken as a turning hypersurface.

in the generalized de Sitter spacetimes is

$$\frac{N_0^{(1)}}{N_0^{(2)}} = \exp\left[\frac{A^{(1)}}{4} - \frac{A^{(2)}}{4}\right]. \tag{15}$$

Integrating the constraint equation $\mathcal{H} = 0$ over a static domain of the Minkowskian spacetime one gets

$$\frac{1}{16\pi}\int \sqrt{-g}R\, d^3x + H_{matter} + \frac{\mathcal{T} A}{4} = 0 \tag{16}$$

where H_{matter} is the total matter energy. The variation of Eq.(16) yields

$$-\frac{\delta A}{4} = \mathcal{T}^{-1}\delta_\lambda H_{matter} \tag{17}$$

where λ labels the *explicit* dependence of H_{matter} on all other (non gravitational) "external" parameters. Eq.(17) is the differential Killing identity of reference 13.

It now follows from Eq.(15) and (17) that matter configurations with neighbouring energies in a static patch of a generalized de Sitter spacetime would be Boltzmann distributed at the global temperature $\mathcal{T}$ provided our ignorance about wormhole states allows to take them to be equally probable. Thus the temperature of the static patch is $\mathcal{T}$ and therefore Eq.(17) also implies that $A/4$ is (up to an integration constant $C_{deSitter}$) the entropy of spacetime and that the latter is in thermal equilibrium with the surrounding matter. As the entropy must be an intrinsic property of spacetime, not only is equilibrium a consequence of the chosen boundary conditions in field space but the converse is also true: the temperature obtained directly from Eq.(17) with the spacetime entropy identified as $A/4 + C_{deSitter}$ must agree at equilibrium with the thermal distribution generated from the field boundary conditions. This justifies a posteriori the above choice of boundary conditions[†].

The tunneling approach to the horizon entropy and temperature[1,2] used here differs from the analysis based on the Euclidean periodicity of Green's functions[8] in two respects. On the one hand, the present approach yields the thermal spectrum, and then the entropy, from the *backreaction* of the thermal matter on the gravitational field, in contradistinction to the Green's function approach. On the other hand however, the thermal matter considered here is taken in the classical limit while the Green's function method describes genuine quantum radiation. Both methods fall short of a fully consistent quantum treatment of the backreaction. But as stated in the introduction the present approach may uncover from the hidden side of the barrier a density of state building the full entropy. Unfortunately, for the de Sitter spacetimes considered above, the hidden side is a wormhole whose description cannot

[†] up to changes which would not alter the probability ratios in the large N_0 limit.

be achieved in our semi-classical approach. Hence, for de Sitter spacetimes, we do not gain at this stage any information on the integration constant $C_{deSitter}$ which measures the density of states left when the full spacetime reduces to the (planckian) wormhole. As we shall now see the situation appears quite different in the case of black hole geometries.

4. The tunneling entropy of black holes

A Schwarzschild static patch of an eternal black hole of mass m_0 is described by the metric

$$ds^2 = (1 - \frac{2m_0}{r})\,dt^2 - (1 - \frac{2m_0}{r})^{-1}\,dr^2 - r^2\,d\Omega^2. \tag{18}$$

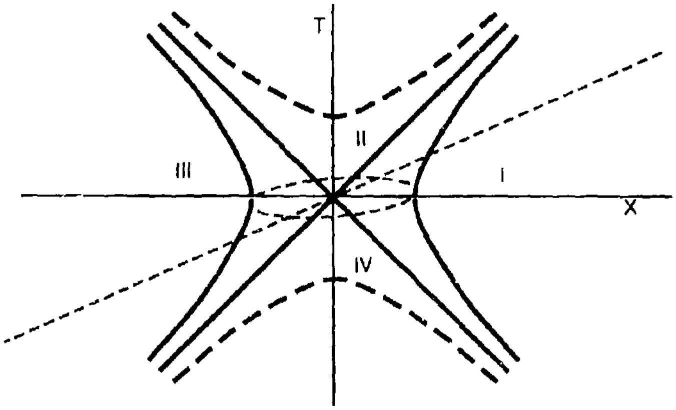


Figure 4. The Kruskal representation of a black hole, eventually surrounded by static matter. The heavy solid line delineates a 4-hyperboloid and the thin one a wormhole. The Euclidean domain $\mathcal{E}$ constituted by a half 4-sphere is delineated by a dashed line. The dotted circle is the turning hypersurface $\tau = 0$.

Surrounding the black hole by static matter generalizes Eq.(18) to

$$ds^2 = g_{00}(r)\,dt^2 - g_{11}(r)\,dr^2 - r^2\,d\Omega^2 \tag{19}$$

where in absence of outer horizon one has

$$r \to \infty : g_{00}(r) = g_{11}^{-1}(r) \to 1 - \frac{2M}{r}. \tag{20}$$

Here $M = M(\infty)$ is the total mass and

$$M(r) = m_0 + \int_{2m_0}^{r} 4\pi z^2 \sigma(z)\, dz$$

$$g_{11}^{-1}(r) = 1 - \frac{2M(r)}{r} \tag{21}$$

$$g_{00}(r) = \left(1 - \frac{2M(r)}{r}\right) \exp\left[-\int_{r}^{\infty} (\sigma + p_1) 8\pi z g_{11}(z)\, dz\right]$$

The metric Eq.(19) can be extended to the four quadrants of a Kruskal space and we choose identical matter distributions in the two Schwartzschild patches to keep a twofold symmetry around the Kruskal time axis. The Kruskal diagram is depicted in Fig.4 where we have also indicated its Euclidean extension $T_e = iT$ resulting from the analytic continuation of the static metric Eq.(19) to the periodic time $t_e = it$.

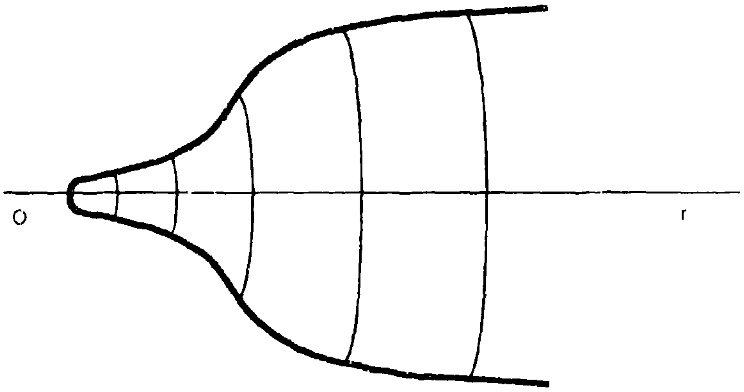


Figure 5. Euclidean black hole surrounded by static matter. Each point is a 2-sphere and the circles span the Euclidean time t_e. The heavy solid line is the turning hypersurface described in Kruskal time by $T = 0$.

The Euclidean periodicity is

$$T = \frac{1}{4\pi}[g_{00}(2m_0)\, g_{11}(2m_0)]^{-1/2} \frac{dg_{00}(r)}{dr}\Big|_{r=2m_0} \tag{22}$$

or from Eq.(21)

$$T = \frac{1}{8\pi m_0} \exp\left[\int_{2m_0}^{\infty} (\sigma + p_1) 4\pi z g_{11}\, dz\right]. \tag{23}$$

The Euclidean extension of the black hole surrounded by static matter is represented in Fig.5. In contradistinction to the de Sitter case, there is clearly no WKB tunneling from a wormhole to a black hole because of the mismatch in topologies. One is therefore lead to investigate possible tunnelings between two black holes geometries $(B.H.)_1$ and $(B.H.)_2$ constituted respectively by black holes of mass m_0 and m $(m > m_0)$ surrounded by matter. The Euclidean sections of $(B.H.)_1$ and $(B.H.)_2$, depicted in Fig.6, are engendered by a rotation of half a Euclidean period of the hypersurfaces a_1 and a_2 labeled by $T = 0$ in their Kruskal diagrams. These are turning hypersurfaces along which Minkowskian and Euclidean black holes meet. We now search for two black holes such that a_1 and a_2 are also the boundaries of an Euclidean solution $\mathcal{E}$ of the Euclidean equations of motion through which tunneling can take place from one Minkowskian black hole to the other. A necessary condition for this to happen is that the total mass $\mathcal{M}$ of the two black hole-matter systems and their Euclidean period $\mathcal{T}^{-1}$ be the same, so that the turning hypersurfaces a_1 and a_2 of the two geometries merge at spatial infinity.

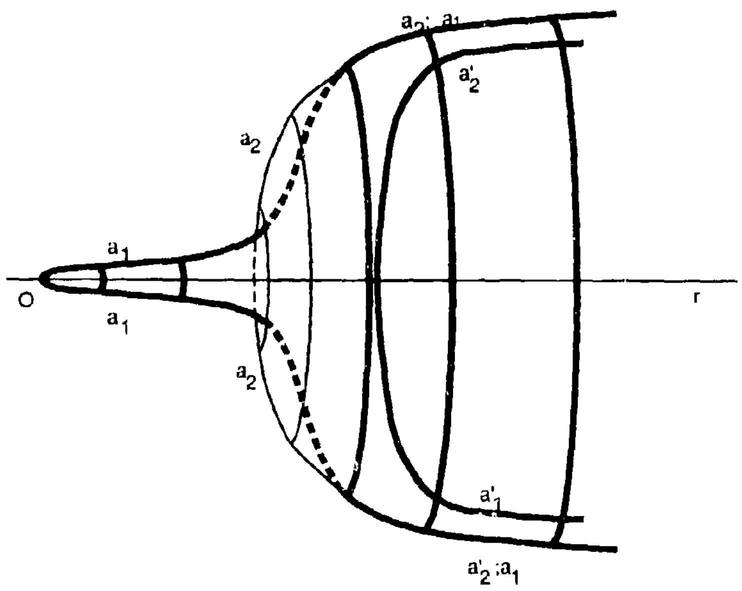


Figure 6. Black hole tunneling. The figure represents the Euclidean sections of the two black hole geometries $(B.H.)_1$ and $(B.H.)_2$. The $(B.H.)_1$ geometry is depicted by thick lines and the $(B.H.)_2$ geometry by thin lines in the region where it differs from the first. The curve a_1 represents a turning hypersurface of $(B.H.)_1$, to be identified with Σ_1. The curve a_2 represents a turning hypersurface of $(B.H.)_2$. The curve a'_2 represents a hypersurface which lays in the intersection of the Euclidean sections of $(B.H.)_1$ and $(B.H.)_2$ and tends to Σ_2 in the limit $m_0 \to 0$.

Let us choose identical matter distributions outside a radius $r_c = 2m + \eta$. η is positive and such that the mass of the matter between the horizon and r_c is 0 for $(B.H.)_2$ and thus $m - m_0$ for $(B.H.)_1$. Keeping m and the matter distribution outside r_c fixed, we now decrease m_0 towards 0. From Eq.(23), in order to keep the Euclidean period T^{-1} constant for $(B.H.)_1$, we have also to decrease η towards 0. As $\eta \rightarrow 0$ the mass $m - m_0$ surrounding the infinitesimal mass m_0 in $(B.H.)_1$ approaches its own Schwartzschild radius. It then follows from Eq.(21) that $g_{00}(r)$ tends to zero for the whole interval $2m_0 < r < 2m$. In other words any frequency stemming from the neighbourhood of the small mass hole is infinitely redshifted by the matter in that interval, as encoded in the damping exponentials in Eq.(21) and (23).

It is clear, from the static coordinate description Eq.(19) extended to Euclidean times $t_e = it$, that the Euclidean sections of $(B.H.)_1$ and $(B.H.)_2$ coincide for $r > 2m + \eta$ but, while for $(B.H.)_2$ the Euclidean section terminates at $r = 2m$, $(B.H.)_1$ presents an extra "needle" in the region $2m_0 < r < 2m$ whose 4-volume is vanishingly small when $\eta \rightarrow 0$. As we now show, this is where tunneling between $(B.H.)_1$ and $(B.H.)_2$ occurs.

To this effect, following the notations of section 2, we identify at finite η the first turning hypersurface Σ_1 through which tunneling takes place with a_1 and consider instead of a second turning hypersurface Σ_2 a hypersurface a_2' which lies in the Euclidean section of both $(B.H.)_1$ and $(B.H.)_2$; thus r is greater than $2m + \eta$ everywhere on a_2'. When $\eta \rightarrow 0$, we can choose a_2' arbitrarily close to a_2. One can then prove[12] that all gravitational momenta vanish in this limit on a_2' in a synchronous system. We may then identify a_2' with Σ_2. The region $\mathcal{E}$ is thus contained in the needle $2m_0 < r < 2m + \eta$. Because of the Kruskal twofold symmetry a_1 is mapped onto itself by a Euclidean time rotation of half a period and thus $\mathcal{E}$ spans only half the needle 4-volume. From Eq.(9), we learn that the inverse transmission coefficient N_0 is simply the exponential of the total Euclidean action of the needle. Although the limiting 4-volume of the needle vanishes, the action will turn out to be finite. It is in fact computable as the difference between the Euclidean actions of the two black holes cut off at the arbitrary radius r_c greater than $2m$ because the two geometries and the two actions coincide for all $r > r_c$.

We thus write

$$N_0 = \exp[S_e^{(B.H.)_2} - S_e^{(B.H.)_1}]. \tag{24}$$

To evaluate these actions we take advantage of the covariance to express them in terms of the static coordinate system Eq.(19) with $t = -it_e$. Using Eq.(10) and (22) and the fact that the integrand in Eq.(10) is the same at r_c for $S_e^{(B.H.)_1}$ and for $S_e^{(B.H.)_2}$, we get

$$N_0 = \exp[4\pi m^2 - 4\pi m_0^2(\eta \rightarrow 0)] \tag{25}$$

or, as m_0 vanishes in the limit,

$$N_0 = \exp A/4 \tag{26}$$

where $A = 16\pi m^2$ is the area of the event horizon of the black hole.

We have thus learned that black holes are related by quantum tunneling to another classical solution for gravity and matter, namely to a "germ black hole" of infinitesimal mass determining the spacetime topology surrounded by a static distribution of matter characterized by a vanishing $g_{00}(r)$. This domain of space-time is characterized by a limiting light-like Killing vector. When the space-time geometry presents a 4-domain endowed with such a Killing vector, we shall call the domain an "achronon". All spherically symmetric achronon configurations will exhibit an infinite time dilation in the Schwarzschild time t, or equivalently massless modes emitted by the achronon are infinitely redschifted. Classically, the achronon has the "frozen" appearance of a collapse at infinite Schwarzschild time. The difference is that it is also frozen in space-time.

To see that achronons can indeed be constructed, at least in a phenomenological fluid model, we shall build a shell model with the required properties.

Let us consider a static spherically symmetric distribution of matter surrounded by an extended shell comprised between two radii r_a and r_b. We define

$$\hat{\sigma} \equiv \int_{r_a}^{r_b} \sigma g_{11}^{1/2} dr, \quad \hat{p}_\theta \equiv \int_{r_a}^{r_b} p_\theta g_{11}^{1/2} dr, \quad \hat{p}_1 \equiv \int_{r_a}^{r_b} p_1 g_{11}^{1/2} dr \tag{27}$$

where $p_\theta = -T^\theta_\theta$ and $p_\phi = -T^\phi_\phi$. Assuming $p_1 = 0$, one may perform the thin shell limit $r_b \to r_a = R$ in these integrals using Eq.(21) and the Bianchi identity

$$p_\theta = p_\phi = \frac{1}{4}(\sigma + p_1)\frac{8\pi r^2 p_1 + 2M(r)/r}{1 - 2M(r)/r} + \frac{1}{2} r p_1' + p_1. \tag{28}$$

One then gets

$$4\pi R\hat{\sigma} = (1 - 2m^-/R)^{\frac{1}{2}} - (1 - 2m/R)^{\frac{1}{2}} \tag{29}$$

$$8\pi R\hat{p}_\theta = \frac{1 - m/R}{(1 - 2m/R)^{\frac{1}{2}}} - \frac{1 - m^-/R}{(1 - 2m^-/R)^{\frac{1}{2}}}; \quad \hat{p}_1 = 0 \tag{30}$$

where m and m^- are the values of $M(r)$ respectively at r_b and r_a and $m_s = m - m^-$ is thus the mass of the shell. Eq.(29) and (30) are the standard result[14]. As the radius R approaches $2m$, these solutions become physically meaningless when $\hat{p}_\theta$ becomes greater than $\hat{\sigma}$: this violates the "dominant energy condition"[15], implying the existence of observers for which the momentum flow of the classical matter becomes spacelike; in fact, the shell is mechanically unstable even before this condition is violated[16].

The divergence of $\hat{p}_\theta$ when $R \to 2m$ appears in Eq.(30) because of the vanishing denominator in Eq.(28). Eq.(30) depends however crucially on the radial pressure being zero inside the shell. Relaxing this condition it is possible to avoid all singularities of the stress tensor as $R \to 2m$ by requiring p_1 inside the shell to satisfy, before performing the thin shell limit,

$$4\pi r^2 p_1 + \frac{M(r)}{r} = 0. \tag{31}$$

This solution is unsatisfactory if the (extended) shell sits in an arbitrary background because of the finite discontinuity of the radial pressure across the shell boundaries which would lead to singularities in p_θ. We may ensure continuity of the radial pressure by immersing the shell in suitable left and right backgrounds. To avoid reintroducing stress divergences when r^b approaches $2M(r^b)$ these should satisfy $(\sigma + p_1) = 0$ at the shell boundaries. One can now perform the thin shell limit. The finite discontinuity of $p_1(r)$ at $r = R$ leads now to

$$\hat{p}_\theta = -\frac{\hat{\sigma}}{2}, \quad \hat{p}_1 = 0 \tag{32}$$

instead of Eq.(30), while $\hat{\sigma}$ is still given by Eq.(29). The dominant energy condition is now satisfied everywhere and provided the background is smooth enough in the neighbourhood of the shell, no stress divergences will appear when it approaches the Schwarzschild radius. Performing the explicit integration over the shell in the exponential term in Eq.(21), one gets for $g_{00}(r)$in the region $0 \leq r < 2m$,

$$g_{00}(r) = (1 - \frac{2M(r)}{r}) \left[\frac{R - 2m}{R - 2m^-}\right] \exp\left[-\mathcal{R}\int_r^\infty (\sigma + p_1) 8\pi z g_{11}\, dz\right]. \tag{33}$$

Here the radius R of the shell is taken at $R = 2m + \eta$ where η is a positive infinitesimal and the symbol $\mathcal{R}$ means that the integral is carried over the regular matter contribution only. Clearly, $g_{00}(r) = O(\eta)$ for $0 \leq r < 2m$, t arbitrary and the above matter distribution constitutes indeed an achronon.

We now relate in general the tunneling as encoded by Eq.(26) to the black hole entropy. This can be done following the analysis of the de Sitter case. Assume all states formed by achronons of mass m surrounded by matter configurations of mass $M - m$, M fixed, to be equally probable. This amounts here to assume the validity of the microcanonical ensemble as achronons can be viewed just as lumps of ordinary matter taken out from the surroundings. The relative probability of finding two black hole geometries for a given total mass M is then given by Eq.(15) with $A^{(1)}$ and $A^{(2)}$ identified here with the black hole areas. The differential Killing identity Eq.(17) follows as before from the integrated constraint equation, the only difference being in general an additional term δM on the right hand side arising from a surface term at spatial infinity. As M is kept fixed, this term plays no role and Eq.(17) remains valid as such. Therefore, in analogy with the de Sitter case, matter configurations with neighbouring energies in a static Schwartszchild patch of an eternal black hole surrounded by matter have a Boltzmann distribution at a global temperature $\mathcal{T}$. The latter now coincides with the local temperature at spatial infinity. $\delta A/4$ is the differential entropy of the hole and $A/4$ is the amount of entropy transferable to matter reversibly. The total black hole entropy is

$$S = A/4 + C_{B.H.} \tag{34}$$

where $\exp C_{B.H.}$ measures the number of quantum states of a residual planckian black hole. The boundary condition in field space at equilibrium are such that the wave functional of an eternal black hole has a small amplitude describing an achronon configuration whose relative weight with respect to the classical black hole configuration is of order $\exp -(A/8)$.

5. From achronon to planckons

We now discuss the nature and the significance of $C_{B.H.}$. The entropy $A/4$ which can be exchanged reversibly from a black hole to ordinary matter was rederived in the preceding section from the existence of a "potential barrier" between a black hole of mass m and an achronon of the same mass. This was done in the context of eternal black holes admitting a Kruskal twofold symmetry, so that there are in fact two achronons imbedded in two causally disconnected static spaces. Within each space black hole-achronon states are in thermal equilibrium with their surroundings. We are therefore led to picture in such a space a quantum black hole eigenstate , in the semi-classical limit, as a quantum superposition of two coherent (normalized) states, $|B.H.\rangle$ and $|A\rangle$ representing respectively a classical black hole and a classical achronon. The relative weight of the two states is approximately, up to a phase, $\exp(-A/8)$. It follows from detailed balance at equilibrium between radiated matter and the black hole that the same superposition should hold for a the black hole who would only emit (and not receive) thermal radiation at the equilibrium temperature. As a black hole formed from collapse indeed emits such a thermal flux we infer that a collapsing black hole is a wave packet formed from a superposition of eigenstates $|C\rangle$ which contain an achronon component with the same weight as in thermal equilibrium. We thus write

$$|C\rangle \simeq |B.H.\rangle + \exp(-A/8)|A\rangle. \tag{35}$$

To a classical single black hole configuration one may associate many distinct classical achronon configurations. In the shell model, for instance, there are infinitely many distinct classical matter configurations of the same total mass m. The argument is however much more general and infinite quantum degeneracy of the achronon is a direct consequence of the infinite time dilation. Indeed, the Hamiltonian H_{matter} is of the form

$$H_{matter} = \int \sqrt{g_{00}} K(\phi_a, g_{ij}, \Pi_a,) \, d^3x \qquad (36)$$

and all its eigenvalues are squashed towards zero by the Schwarzschild time dilation factor $\sqrt{g_{00}}$, thus generating an infinite number of orthogonal zero energy modes on top of the original achronon.

The infinity of zero energy modes around any background implies an infinite degeneracy of achronons of given mass and thus an infinity of distinct quantum black hole states of the same mass differing by the achronon component of their wave function. This infinite degeneracy of the quantum black hole provides the reservoir from which are taken the finite number of "surface" quantum states $\exp A/4$ counted by the area entropy $A/4$ transferable reversibly to outside matter.

Except for providing a rational for the large but finite testable entropy of the black hole, achronons do not modify the behaviour of large macroscopic black holes. However when their mass is reduced by evaporation and approaches the Planck mass the barrier disappears and quantum superposition completely mixes the two components. Of course, this means that both the description in terms of semiclassical configurations and of tunneling disappears. What remains however as a consequence of unitarity, is the infinity of distinct orthogonal quantum states available which have no counterpart in the finite number of decayed states. The quantum black hole has become a planckon[12], that is a Planckian mass object with infinite degeneracy. In terms of Eq.(34), this means that in a asymptotically flat background, the integration constant of the black hole entropy $C_{B.H.}$ is infinite. As discussed in the introduction, this means that in such a background a generic planckon cannot decay nor be formed in a finite time.

This conclusion however is contingent upon the validity, at the qualitative level, of our semi classical approach. The main question is whether or not the *quantum* backreaction of the matter on the metric removes the infinite degeneracy. Answering it requires further analysis.

Finally, it is of interest to note that the planckon solution to the unitarity problem posed by the evaporating black hole would have, at a fundamental level, far reaching implications on the spectrum of quantum gravity. The opening at the Planck size of an infinite number of states, an unavoidable consequence of the existence of planckons, may appear as a horrendous complication which could make quantum gravity definitely unmanageable but hopefully the converse may be true. Indeed planckons should make quantum gravity ultraviolet finite. The Hilbert space of physical states available to macroscopic observer must be orthogonal to the infinite set of states describing Planckian bound states. Their wave function at Planckian scales where planckon configurations are concentrated are therefore expected to be vanishingly small. In this way, planckons would provide the required short distance cut-off for a consistent field theoretic description of quantum gravity within our universe while leaving the largest part of its information content hidden at the Planck scale.

An operational formulation of quantum gravity applicable within our universe and based on conventional four dimensional gravity might thus well be within reach. Nevertheless, the sudden widening of the spectrum of physical states at the Planck scale and the relative scarcity of states which describe large distance physics suggest that a fully consistent theory cannot be formulated in terms of only long range quantum fields (including the metric), and a larger scheme may be required to cope with the infinite amount of information relegated to the Planck scale.

Acknowledgements

We are very grateful to R. Balbinot, R. Brout, J. Katz, J. Orloff, R. Parentani and Ph. Spindel for most enjoyable, stimulating and clarifying discussions.

References

1. F. Englert,"From Quantum Correlations to Time and Entropy" in *The Gardener of Eden; Physicalia Magazine* (special issue in honour of R. Brout's birthday), (1990) Belgium Ed. by P. Nicoletopoulos and J. Orloff.
2. A. Casher and F. Englert, *Class. Quantum Grav.***9** (1992) 2231.
3. B. De Witt, *Phys. Rev.* **160** (1967) 113.
4. T. Regge and C. Teitelboim, *Annals of Physics* **88** (1974) 286.
5. D. Page and K.W. Wooters, *Phys. Rev.* **D27** (1983) 2885.
   Y. Aharonov and T. Kaufherr, Phys. Rev. **D30** (1984) 368.
   J.B. Hartle, Phys. Rev. **D37** (1988) 2818; Phys. Rev. **D38** (1988) 2985
6. T. Banks, *Nucl. Phys.* **249** (1985) 332,
   R. Brout, *Foundations of Physics* **17** (1987) 603
   R. Brout, G. Horwitz and D. Weil,*Phys. Lett.* **B192** (1987) 318
7. W. Unruh and W.H. Zurek, *Phys. Rev.* **D40** (1989) 1064.
   J.J. Halliwell,*Phys. Rev.* **D39** (1989) 2912.
8. G. Gibbons and S. Hawking, *Phys. Rev.* **D15** (1977) 2738; 2752.
9. Y. Aharonov, A. Casher and S. Nussinov, *Phys. Lett.* **B191** (1987) 51
10. J.A. Harvey and A. Strominger, *Quantum Aspects of Black Holes*, Preprint EFI-92-41; hep-th/9209055.
11. T. Banks, M O'Loughlin and A .Strominger , *Black Hole Remnants and the Information Puzzle*, Preprint RU-92-40; hep-th/9211030.
12. A. Casher and F. Englert, *Black Hole Tunneling Entropy and the Spectrum of Gravity*, Preprint ULB TH 8/92, TAUP 2017-92; gr-qc/9212010.
13. J M. Bardeen, B. Carter and S.W. Hawking, *Comm. Math. Phys.* **31** (1973) 161

14. J. Frauendiener, C. Hoenselaers and W. Conrad, *Class. Quantum Grav.* 7 (1990) 585.

15. S.W. Hawking and G.F.R. Ellis, *The Large Scale of Space-Time*, (Cambridge University Press, Cambridge, England, 1973).

16. P.R. Brady, J Louko and E. Poisson, *Phys. Rev.* **D44** (1991) 1891.

Black Holes and Information Loss in 2D Dilaton Gravity

S. P. de Alwis
Department of Physics, Campus Box 390
University of Colorado
Boulder, CO 80309, USA

ABSTRACT

The theory of dilaton gravity coupled to conformal matter proposed by Callan *et al.* is quantized. An attempt is made to interpret the resulting picture of quantum effects in terms of Hawking radiation. The question of information loss is also addressed briefly.

There are three main responses to Hawking's original observation[1] that the semi-classical calculation of quantum effects around a black hole, led to its evaporation via thermal radiation, and the corresponding loss of information that went into the hole.

a) Pure states evolve into mixed states. This is Hawking's position. [2]

b) The black hole does not evaporate completely. A remnant which stores all the information is left behind. This is the position advocated by Aharanov, Casher and Nussinov.[3]

c) The radiation is not exactly thermal. There are subtle correlations which code the information. This is the position first advocated by Page[4].

The first of these entails a radical reformulation of quantum mechanics and in particular one looses the connection between symmetry and conservation laws.[5] One should probably show that more conservative options are completely ruled out before accepting such a radical revision of the foundations of quantum mechanics. The second has the problem that it leads to a an infinite degeneracy of states. The third alternative is the most conservative but it is yet to be demonstrated even in a toy model.

The theory of dilaton gravity coupled to conformal matter proposed by Callan, Giddings, Harvey, and Strominger[6] (CGHS) is in fact a toy model in two dimensions within which one may hope to gain some understanding of these issues. In this talk I'm going to discuss the quantization of the CGHS theory and its physical implications. This talk is based on the papers of reference 7. Similar work has been done by Bilal and Callan[8]. (The discussion of ADM and Bondi masses in the last few paragraphs of this paper has been revised in accordance with reference 11.)

The CGHS action is given by

$$S = \frac{1}{4\pi} \int d^2\sigma \sqrt{-g} [e^{-2\phi}(R + 4(\nabla\phi)^2 + 4\lambda^2) - \frac{1}{2}\sum_{i=1}^{N}(\nabla f^i)^2]. \tag{1}$$

In the above G is the $2d$ metric, R is its curvature scalar, ϕ is the dilaton and the f^i are N scalar matter fields.

The quantum field theory of this classical action may be defined as

$$Z = \int \frac{[dg]_g [d\phi]_g [df]_g}{[Vol.\ Diff.]} e^{iS[g,\phi,f]}. \tag{2}$$

Now let us gauge fix to the conformal gauge $g = e^{2\rho}\hat{g}$ and rewrite the measures with respect to the fiducial metric $\hat{g}$. Following the work of David and of Distler and Kawai[9], we may expect the action to get renormalized, except that unlike in their case the renormalization will be dilaton dependent (since the coupling is $e^{2\phi}$). Thus in general we may expect the gauge fixed path integral to be written as

$$Z = \int [dX^\mu]_{\hat{g}} [df]_{\hat{g}} ([db][dc])_{\hat{g}} e^{iI(X,\hat{g}) + iS(f,\hat{g}) + iS(b,c,\hat{g})}, \tag{3}$$

where

$$I[X,\hat{g}] = -\frac{1}{4\pi} \int \sqrt{-\hat{g}} [\frac{1}{2} \hat{g}^{ab} G_{\mu\nu} \partial_a X^\mu \partial_b X^\nu + \hat{R}\Phi(X) + T(X)]. \tag{4}$$

$S(b,c,\hat{g})$ is the Fadeev-Popov ghost action, and we have written $(\phi,\rho) = X^\mu$. Note that all the measures in Eq. (3) are defined with respect to the $2d$ metric $\hat{g}$ and that in particular the measure $[dX^\mu]$ is derived from the natural metric on the space $||\delta X_\mu||^2 = \int d^2\sigma \sqrt{-\hat{g}} G_{\mu\nu} \delta X^\mu \delta X^\nu$.

The only a priori restriction arises from the fact that the functional integral for Z in Eq. (3), must be independent of the fiducial metric $\hat{g}$, as is obvious from the expression Eq. (2) for it. i.e. we must have

$$< T_{\pm\pm} + t_{\pm\pm} >= 0, \tag{5}$$

and $< T_{+-} + t_{+-} >= 0$. The latter is equivalent to the ρ equation of motion. In addition we must have the integrability conditions for the above constraints, i.e. that they generate a Virasoro algebra with zero central charge. This requirement is equivalent to the condition that the beta-function equations for G, Φ, and T are satisfied. Thus we must have,

$$\begin{aligned} \beta_{\mu\nu} &= \mathcal{R}_{\mu\nu} + 2\nabla^G_\mu \partial_\nu \Phi + \ldots, & (6) \\ \beta_\Phi &= -\mathcal{R} + 4G^{\mu\nu}\partial_\mu\Phi\partial_\nu\Phi - 4\nabla^2_G\Phi + \frac{(N+2)-26}{3} + \ldots, & (7) \\ \beta_T &= -2\nabla^2_G T + 4G^{\mu\nu}\partial_\mu\Phi\partial_\nu T - 4T + \ldots, & (8) \end{aligned}$$

where R is the curvature of the metric G. These equations have to be solved under the boundary conditions that in the weak coupling limit ($e^{2\phi} << 1$) we get, (comparing Eq. (4) with the classical CGHS action in the conformal gauge with the conformal anomaly term added)

$$G_{\phi\phi} = -8e^{-2\phi},\ G_{\phi\rho} = 4e^{-2\phi},\ G_{\rho\rho} = 2\kappa,\ \Phi = -e^{-2\phi} + \kappa\rho,\ T = -4\lambda^2 e^{2(\rho-\phi)}. \tag{9}$$

The (renormalized) field space metric may be parametrized as,

$$ds^2 = -8e^{-2\phi}(1 + h(\phi))d\phi^2 + 8e^{-2\phi}(1 + \overline{h}(\phi))d\rho d\phi + 2\kappa(1 + \overline{\overline{h}})d\rho^2, \qquad (10)$$

where $h, \overline{h}$, and $\overline{\overline{h}}$ are $O(e^{2\phi})$. If we are going to consider only $O(e^{2\phi})$ effects then we should certainly set $\overline{\overline{h}}$ to zero. But even if we consider the renormalization functions h and $\overline{h}$ to all orders, it is consistent to limit ourselves to the class of quantum versions of the CGHS theory which have $\overline{\overline{h}} = 0$, provided that we satisfy the beta function equations. This corresponds to confining ourselves to theories in which the field space curvature $\mathcal{R} = 0$. In this case we can transform this metric to Minkowski form. Putting $y = \rho - \kappa^{-1}e^{-2\phi} + \frac{2}{\kappa}\int d\phi e^{-2\phi}\overline{h}(\phi)$ and $x = \int d\phi P(\phi)$, $P(\phi) = e^{-2\phi}[(1+\overline{h})^2 + \kappa e^{2\phi}(1+h)]^{\frac{1}{2}}$, we have $ds^2 = -\frac{8}{\kappa}dx^2 + 2\kappa dy^2$. Then demanding that we recover the CGHS Φ given in Eq. (9) in the weak coupling limit we find the unique solution $\Phi = \kappa y$ from the first beta function equation, and substituting in the second beta function equation we get $\kappa = \frac{24-N}{6}$. The third beta function equation together with the boundary condition that we get the CGHS value for T in the weak coupling limit then gives $T = -4\lambda^2 e^{-\frac{4}{\kappa}x+2y}$.

Introducing rescaled fields, $X = 2\sqrt{\frac{2}{|\kappa|}}x$, $Y = \sqrt{2|\kappa|}y$, we then have the functional integral,

$$Z = \int [dX][dY][df][db][dc]e^{iS[X,Y,f]+iS_{ghost}}, \qquad (11)$$

where,

$$S = \frac{1}{4\pi}\int d^2\sigma[\mp\partial_+X\partial_-X \pm \partial_+Y\partial_-Y + \sum_i \partial_+f^i\partial_-f^i + 2\lambda^2 e^{\mp\sqrt{\frac{2}{|\kappa|}}(X\mp Y)}]. \qquad (12)$$

Now the transformation from ϕ to X is singular if $P(\phi)$ has a zero.But there is a whole class of functions h and $\overline{h}$ for which this is not the case (the simplest example being $h = 0, \overline{h} == -\frac{\kappa}{2}e^{2\phi}$ so that in Eq. 11) the integration range goes over the whole real line and we have (by slightly generalizing the arguments given in reference 10) an exact conformal field theory as is required by the general covariance of the original theory. The equations of motion for X, Y coming from Eq. (12) can be solved in terms of four arbitrary chiral functions. Indeed they are the same solutions as the classical CGHS ones the quantum anomalies being hidden in the relation between X, Y and ϕ, ρ. By a coordinate choice two of the functions can be set to zero so that in this conformal frame(which I will call the Kruskal frame) one has $X = -Y = -\sqrt{\frac{2}{|\kappa|}}(u - \lambda^2\sigma^+\sigma^-)$ where $u = u(\sigma^+) + u_-(\sigma^-)$. To be explicit consider the case d) discussed at the end of the last section ($h = 0$, $\bar{h} = -\frac{\kappa}{2}e^{2\phi}$); then $X = 2\sqrt{\frac{2}{|\kappa|}}\int d\phi e^{-2\phi}[1+\frac{\kappa^2}{4}e^{4\phi}]^{\frac{1}{2}} = \sqrt{2|\kappa|}\int d\phi[1+\frac{4}{\kappa^2}e^{-4\phi}]^{\frac{1}{2}}$, and $Y = \sqrt{2|\kappa|}\rho + \sqrt{\frac{2}{|\kappa|}}e^{-2\phi} - \sqrt{2|\kappa|}\phi$.

In the weak coupling limit ($e^{2\phi} << 1$) we have from the solution for X, Y and the above, the classical solution $e^{-2\phi} = e^{-2\rho} = u - \lambda^2\sigma^+\sigma^-$, which exhibits the classical (black hole type) singularity on the curve where the right hand side vanishes. But the singularity is in the strong coupling region where we have to use the strong coupling expansion (from the second line of the above equation for X) $X \simeq \sqrt{2|\kappa|}[\phi - \frac{e^{-4\phi}}{\kappa^2} + \ldots]$.

Then we have from the solution for X,Y, $\phi \simeq \kappa^{-1}(u - \lambda^2\sigma^+\sigma^-)$, and $\rho \simeq \frac{1}{\kappa}e^{-2\kappa^{-1}(u-\lambda^2\sigma^+\sigma^-)}$. The metric ($e^{2\rho}$) is clearly non-singular at the classical singularity.

The $u_\pm$ are fixed in terms of the ghost and f-matter stress tensors by the constraints Eq. (5), (for a detailed discussion see the third paper of reference 2). To proceed one has to make an assumption about these stress tensors. In order to be as close as possible to the original Hawking calculation[6], I assume that in a preferred coordinate system which is asymptotically Minkowski (which is the natural frame to choose at null past infinity $\mathcal{I}^+_R$) the expectation value of the matter stress tensor is zero. This still leaves an ambiguity in the ghost influx. Choosing the latter to be constant and equal for both left and right movers, and transforming to the Kruskal frame by taking into account the conformal anomaly (Schwartz derivative term), and using the constraint equations, we get $u_+ = a_+ + b_+\sigma^+ - a(\sigma^+ - \sigma^+_0)\theta(\sigma^+ - \sigma^+_0) - \frac{\bar{N}}{24}\log|\sigma^+|$, $u_- = a_- + b_-\sigma^- - \frac{\bar{N}}{24}\log|\sigma^-|$, where $\bar{N} = N + \alpha - 26$. Here α is an arbitrary parameter characterizing the ghost influx. Perhaps the most natural choice is to put $\alpha = 26$ and we shall do so in the following. With $a = b = 0$ one has the quantum analog of the static black hole solutions of dilaton gravity. The step function term comes from assuming (as in CGHS[6]) that the matter falls in the form of a shock wave. It is trivial to generalize this to more general configurations of infalling matter. So $a_\pm = b = 0$, $a \neq 0$ corresponds to a dynamical solution describing collapse to a black hole and Hawking radiation.

One may now derive[12] an expression for the ADM and Bondi masses of these configurations following an argument of Regge and Teitelboim[13].

$$E_{ADM} = \sqrt{\frac{|\kappa|}{2}}[\frac{1}{2}g'(\sigma)\Delta Y - \Delta Y']^{\infty}_{-\infty}. \tag{13}$$

In the above g reflects the freedom in the choice of conformal frame and is zero in the Kruskal frame.

For the static solutions one finds from this that the ADM mass is zero. In the case of the dynamic solutions one gets an ADM mass in these coordinates that is equal to the energy of the infalling matter $M_0 = \lambda a\sigma^+_0$. Thus the ADM mass in fact satisfies a positive energy theorem. However the frame appropriate to an observer outside the horizon is related to the frame σ by $\overline{\sigma}^+ = \frac{1}{\lambda}e^{\lambda\overline{\sigma}^+}, \sigma^- = -\frac{1}{\lambda}e^{-\lambda\overline{\sigma}^-} - \frac{a}{\lambda^2}$. In this frame one gets for the dynamical solution,

$$E_{ADM} = M_0 + \frac{N}{24}\ln\left(1 + \frac{a}{\lambda}e^{\lambda(\overline{\tau}-\overline{\sigma})}\right)|_{\overline{\sigma}\to-\infty} + \frac{N}{24}\lambda. \tag{14}$$

Thus we have an infinite value for the energy in these coordinates. It should be noted that the corresponding classical solution has a finite mass equal to the incoming matter energy. Thus the infinite value is a consequence of the quantum radiation.

One may now calculate the Bondi mass (i.e. the mass left over after radiation for a light cone time σ^-) to get

$$\begin{aligned} E_{Bondi}(\overline{\sigma}) &= M_0 - \frac{N}{24}\ln(1+\frac{a}{\lambda}e^{\lambda\overline{\sigma}^-}) - \frac{N}{24}\frac{\lambda}{(1+\frac{\lambda}{a}e^{-\lambda\overline{\sigma}^-})} \\ &- \left[-\frac{N}{24}\ln(1+\frac{a}{\lambda}e^{\lambda\overline{\sigma}^-}) - \frac{N}{24}\frac{\lambda}{(1+\frac{\lambda}{a}e^{-\lambda\overline{\sigma}^-})}\right]_{\overline{\sigma}^-\to\infty} \end{aligned} \qquad (15)$$

For any finite value of $\overline{\sigma}^-$ one gets an infinite value reflecting the fact that the ADM mass is infinite. However at $\overline{\sigma}^- \to \infty$ one finds that the energy left behind is equal to the incoming matter energy M_0. This somewhat peculiar conclusion seems to be forced upon us by the conformal invariance of the two dimensional theory. Since the collapsing mass M_0 is left behind it might seem that the model supports the remnant scenario[3]. However this is the total mass of the original collapsing matter and the bath of radiation was there ab initio. Thus it is not really possible to draw any conclusion for a situation that one might obtain in four dimensions, where one is not forced by conformal invariance to start with an infinite bath of radiation.

Acknowledgements

This work was partially supported by Department of Energy contract No. DE-FG02-91-ER-40672.

7. References

1. S.W. Hawking.*Comm. Math. Phys.* **43** (1975) 199.
2. S. W. Hawking, *Comm. Math. Phys.* **87** (1982) 395.
3. Y. Aharanov, A. Casher, and S. Nussinov, *Phys. Lett.* **B191** (1987) 51.
4. D. Page, *Phys. Rev. Lett.* **44** (1980) 301.
5. T. Banks, M. Peskin, L. Susskind, *Nucl. Phys.* **B244** (1984) 125.
6. C.G. Callan, S.B. Giddings, J.A. Harvey, and A. Strominger, *Phys. Rev.* D45 (1992) R1005.
7. S. P. de Alwis, *Phys. Lett.* **B289** (1992) 278; Colorado preprint, (1992) COLO-HEP-284, hep-th/9206020 *Phys. Lett.* **B** in press, and *Phys. Rev.* **D15**, (1992) 5429.
8. A. Bilal and C. Callan, Princeton preprint (1992) PUPT-1320, hepth@xxx/9205089.
9. F. David, *Mod. Phys. Lett.* **A3** (1988) 1651, J.Distler and H. Kawai, *Nucl. Phys.* **B321** (1989)509.
10. T. Curtright and C. Thorn, *Phys. Rev. Lett.* **48** (1982) 1309; E. Braaten, T. Curtright, and C. Thorn, *Phys. Lett.* **B118** (1982) 115; *Ann. Phys.* **147** (1983) 365; E. D'Hoker and R. Jackiw, *Phys. Rev.* **D26** (1982) 3517.
11. S. P. de Alwis, Colorado preprint COLO-HEP-309, hep-th/9302144.
12. T. Regge and C. Teitelboim *Ann. Phys.* **88** 286 (1974).

SECTION 5

GRAVITATION AND QUANTUM MECHANICS

TOPOLOGICAL PHASES AND THEIR DUALITY IN ELECTROMAGNETIC AND GRAVITATIONAL FIELDS

J. ANANDAN
Department of Physics and Astronomy, University of South Carolina
Columbia, SC 29208, USA

ABSTRACT

The duality found by Aharonov and Casher for topological phases in the electromagnetic field is generalized to an arbitrary linear interaction. This provides a heuristic principle for obtaining a new solution of the field equations from a known solution. This is applied to the general relativistic Sagnac phase shift due to the gravitational field in the interference of mass or energy around a line source that has angular momentum and the dual phase shift in the interference of a spin around a line mass. These topological phases are treated both in the linearized limit of general relativity and the exact solutions for which the gravitational sources are cosmic strings containing torsion and curvature, which do not have a Newtonian limit.

0. Introduction

As is well known, some of Yakir Aharonov's most famous contributions concern topological phases due to the electromagnetic field. It is therefore fitting on this occasion of his sixtieth birthday for me to present him with some observations concerning these phases, which generalize naturally to the gravitational field. In particular, I shall examine the duality between the Aharonov-Bohm (AB) phase [1] and the phase shift in the interference of a magnetic moment in an electric field [2] which was found by Aharonov and Casher (AC) [3]. I shall show, by means of the linearized limit and an exact solution of the gravitational field equations, that both these phases have gravitational analogs and they satisfy this duality.

In section 1, I shall briefly summarize the phase shifts in the interference of a charge and a magnetic dipole (at low energies) due to the electromagnetic field. These phase shifts reveal, respectively, $U(1)$ and $SU(2)$ gauge field aspects of the electromagnetic field. But these two aspects are not independent: The $SU(2)$ connection which gives the dipole phase shift depends on the electric and magnetic fields and as such are derived from the electromagnetic connection that gives the $U(1)$ AB phase shift. It is nevertheless amusing to see a charged particle with a magnetic moment, such as an electron, interacting with an electromagnetic field as if it is a $U(1) \times SU(2)$ gauge field. Two topological phase shifts due to electric and manetic fields corresponding to two $U(1)$ subgroups of $SU(2)$ will be reviewed. The duality between one of these and the AB phase, found by AC, will be generalized to an arbitrary interaction in section 2. I shall formulate a duality principle which states that any two dual phases are equal under certain conditions.

The gravitational phase shifts, obtained in section 3, are special cases of the phase shifts obtained previously [4, 5] due to the coupling of the mass and spin to the gravitational field. The key to the analogy with the electromagnetic phase shifts is that the mass or energy plays the role of the electric charge and spin the role of the magnetic dipole in the electromagnetic field. The gravitational phase shifts are the same as due to the phase shifts of a Poincare gauge field. The translational and Lorentz aspects of the Poincare group are respectively analogous to the $U(1)$ and $SU(2)$ aspects of the electromagnetic field, mentioned above.

If gravity contains torsion, as will be assumed here, the connection, which gives the phase shift of the spin, is independent of the metric or the vierbein, which gives the phase shift due to the mass or energy-momentum. Therefore, these two aspects are then complementary, unlike in the electromagnetic case in which the $SU(2)$ connection depends on the $U(1)$ connection as mentioned earlier. The electromagnetic field and its sources of course must satisfy the Maxwell's equations. It is well known that the solenoid which produces the AB phase shift is a solution. Similarly, the gravitational field and its sources must satisfy Einstein's field equations or a suitable generalization of it to include torsion. Fortunately, an exact solution corresponding to a spinning cosmic string with angular momentum and mass, which is the analog of the solenoid with a coaxial line charge in the electromagnetic case, can be obtained everywhere including the interior of the string.

There is a topological general relativistic Sagnac phase [4] which depends on the energy of a particle outside the string and the flux of torsion inside the string produced by its spin. This is analogous to the AB phase. There is another topological phase which depends on the spin of the particle and the flux of curvature inside the string, produced by its mass. This is the dual of the former phase. I shall show that this pair of dual phases satisfy the duality principle formulated in section 2.

1. Topological and Geometrical Phases Due to the Electromagnetic Field

For simplicity, consider the non relativistic Hamiltonian of a charged particle in an electromagnetic field

$$H = \frac{1}{2m}(\mathbf{p} - e\mathbf{A})^2 + eA_o, \tag{1.1}$$

where e and m are the charge and mass of the particle and $A_\mu = (A_0, A_i) = (A_0, -A^i)$ is the electromagnetic potential representing the $U(1)$ connection due to this field*.

* Units in which the velocity of light $c = 1$ will be used throughout.

As is well known, (1.1) predicts the Aharonov- Bohm (AB) effect [1]. This is the electromagnetic phase difference between two interfering coherent beams which are entirely in a multiply connected region in which the field strength $F_{\mu\nu}$ is zero. The phase factor that determines the electromagnetic shift in the interfering fringes is

$$\Phi_C = \exp(-\frac{ie}{\hbar}\oint_C A_\mu dx^\mu), \qquad (1.2)$$

where the closed curve C passes through the two interfering wave functions, and encloses a region in which the field strength is non zero. Consequently, this phase shift is constant as the beams are varied in the outside region in which the field strength is zero, which makes this effect topological. Conversely, the experimental observation of this phase shift may be used to infer that the electromagnetic field is a $U(1)$ gauge field [6,7].

Consider now the interaction of a neutral magnetic dipole, such as a neutron, with the electromagnetic field which at low energies is described by the Hamiltonian [2,3,8,9]

$$H = \frac{1}{2m}(\mathrm{p} - \gamma \mathbf{A}^k S_k)^2 + \gamma A_o{}^k S_k, \qquad (1.3)$$

where $A_\mu{}^k = (A_0{}^k, A_i{}^k) = (-B^k, \epsilon_{ijk}E^j)$ in terms of the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$, $S_k, k = 1,2,3$ are the spin components which generate the $SU(2)$ spin group. For a spin $\frac{1}{2}$ particle, the magnetic moment $\mu = \frac{\gamma\hbar}{2}$. This interaction is like that of an isospin with an $SU(2)$ Yang-Mills field [8,9,10].

The phase shift due to both electric and magnetic fields, in the interference of a neutral dipole such as a neutron, was obtained by means of an explicit plane wave solution [2]. This result, of course, applies also to the more general situation when the interfering wave functions are locally approximate plane waves so that the WKB approximation is valid. Hence, the phase shift is determined by the non abelian phase factor

$$\Phi_C = P\exp(-\frac{i\gamma}{\hbar}\oint_C A_\mu{}^k S_k dx^\mu), \qquad (1.4)$$

where P denotes path ordering, and C is a closed curve consisting of unperturbed trajectory [8]. Hence, Φ_C is an element of $SU(2)$, and this phase shift is like the phase shift due to an $SU(2)$ gauge field [5,6,11].

The special case when the two waves interfere around a line charge was considered by AC [3]. In this case, $A_0{}^k = 0$ and the electric field E^j and therefore $A_i{}^k = \epsilon_{ijk}E^j$ fall off inversely as the distance from the line charge. It follows immediately from (1.4) that, if the spin is polarized parallel to the line charge, then this phase shift is topological in the sense that it does not change when the curve C surrounding the line charge is deformed. However, the Yang-Mills field strength $F^k_{\mu\nu}$ of $A_\mu{}^k$ is non vanishing outside the line charge, which makes this effect fundamentally different from the AB effect in which the electromagnetic field strength $F_{\mu\nu} = 0$ along the beams. But if the line charge is in the 3-direction then $F^3_{\mu\nu} = 0$. That is the field strength corresponding to the $U(1)$ subgroup of the spin $SU(2)$ generated by S_3 is zero. So, for this subgroup, this phase shift is like the AB effect.

Another topological phase shift experienced by the dipole is the following: The wave packet of a neutron or an atom is split into two coherent wave packets, one of which enters a cylindrical solenoid. The homogeneous magnetic field of the solenoid is then turned on and is then turned off before the wave packet leaves the solenoid. Then there is a phase shift even though there is no force acting on the neutron. This phase shift, which is easily obtained from (1.4), is due to $A_0{}^k = -B^k$. Hence, this phase shift is due to the scalar potential of the gauge field of the $U(1)$ subgroup of the spin $SU(2)$ group generated by the component of $\mathbf{S}$ in the direction of $\mathbf{B}$. At the suggestion of Zeilinger [12] and the author [8] this experiment was performed for neutrons [13].

The general case of the phase shifts for a particle that has charge and magnetic moment interacting with an electromagnetic field was studied before [8,9]. I shall restrict here, for simplicity, the special case of the particle being a Dirac electron with "g-factor" being two. Its Hamiltonian, at low energies in the inertial frame of the laboratory, is

$$H = \frac{1}{2m}(\mathbf{p} - e\mathbf{A} - \frac{1}{2}\gamma' \mathbf{A}^k S_k)^2 + eA_o + \gamma' A_o{}^k S_k, \tag{1.5}$$

This Hamiltonian is like as if the electron is interacting with a $SU(2) \times U(1)$ gauge field represented by the gauge potentials $\mathbf{A}^k$ and $\mathbf{A}$. Note, however, the factor $\frac{1}{2}$ in front of $\mathbf{A}^k$ compared to (1.3). This is due to the Thomas precession undergone by the electron when it accelerates in the electric field [2,9].

2. The Duality of AC and Its Generalization

The major new contribution of AC, which is not contained in any earlier work, is the recognition that the phase shift due to the line charge is "dual" to the AB effect due to a solenoid. I shall now give a precise statement of this duality which would be general enough to apply to other interactions as well.

Suppose that an infinite uniform solenoid is situated along the z-axis of a Cartesian coordinate system. A charge of strength e is taken slowly around the solenoid along a circle in the xy-plane with its center at the solenoid, which is assumed to have negligible cross-section. The solenoid may be regarded as a magnetized medium with a constant magnetic moment per unit length equal to $\mathbf{M}$, say, which is parallel to the z- axis. The AB phase shift acquired by the charge

$$\Delta\phi = \frac{e}{\hbar}\int_{\Sigma} \mathbf{B} \cdot d\mathbf{\Sigma} = \frac{eM}{\hbar}, \tag{2.1}$$

where Σ is a cross-section of the solenoid, $\mathbf{B}$ is the magnetic field inside the solenoid, and $M = |\mathbf{M}|$.

Now, divide the solenoid into slices each of height $\delta\ell$ bounded by cross-sections that are parallel to the xy-plane. The magnetic moment of each slice is

$$\mu = \mathrm{M}\delta\ell. \tag{2.2}$$

The linearity of Maxwell's equations imply that the phase shift is *linear* in the sense that (2.1) is the sum of the phase shifts due to the influence of each slice of the solenoid on the charge. Consider a slice whose center is at $z = Z$. Then taking the charge around the circle mentioned above is equivalent to keeping the charge fixed at $z = -Z$ and taking the slice of magnetic moment $\mu = |\mu|$ around the same circle in the xy- plane. The phase shifts acquired in both processes are the same. This has been shown using space-time translation and Galilei invariance of the Lagrangian, for the special case of charge-dipole interaction [3] and using Lorent invariance for the general case of an arbitrary interaction [9]. Now do this for each *pairwise* interaction between the charge e and each slice with magnetic moment μ. Then, as we account for all slices from $z = +\infty$ to $z = -\infty$, in the new situation, which will be called the dual of the original situation, there are charges from $z = -\infty$ to $z = +\infty$ along the z-axis, and the magnetic moment μ circles around this line charge. Each charge e is contained in an interval of height $\delta\ell$, and may be assumed to be spread uniformly in that interval. Therefore,

$$e = \lambda\delta\ell, \tag{2.3}$$

where λ is the charge per unit length. It follows that the magnetic moment which circles around this line charge, with its direction parallel to the z-axis, acquires a phase shift equal to $\Delta\phi$ given by (2.1). From (2.2) and (2.3),

$$\frac{e}{\lambda} = \frac{d}{M}. \tag{2.4}$$

for these two dual situations. Using (2.2), (2.1) may be rewritten as

$$\Delta\phi = \frac{\lambda\mu}{\hbar}. \tag{2.5}$$

This phase shift may also be independently derived using (1.4) and the electric field of a line charge obtained by solving Maxwell's equations.

The above argument may be generalized to the case when the charge goes around an arbitrary closed curve $\mathbf{r}(t)$ which may or may not enclose the solenoid. Then relative to one of the above mentioned slices at say $\mathbf{Z} = (0, 0, Z)$ this curve is $\mathbf{r}(t) - \mathbf{Z}$. Therefore, in the dual situation the slice with magnetic moment μ moves around the closed curve $-\mathbf{r}(t) + \mathbf{Z}$ relative to the charge. So, if the charge is placed at $-\mathbf{Z}$, the slice goes around the closed curve $-\mathbf{r}(t)$. By doing this for each pairwise interaction between the charge and the fixed magnetic moment of each of the slices into which the solenoid is divided, I obtain the dual situation in which a magnetic dipole of strength μ, and direction parallel to the z-axis, moves around the closed curve $-\mathbf{r}(t)$ with a line charge, whose charge per unit length is λ, along the z-axis. Also, since this interaction is invariant under parity, the same phase shift is obtained for the situation obtained by parity transforming about the origin. This corresponds to the magnetic moment moving around the original curve $\mathbf{r}(t)$ traveled by the charge when in the presence of the solenoid.

Now the statement that the AB phase shift is topological may be expressed as follows: If the curve $\mathbf{r}(t)$ goes around the solenoid n times, $n = 0, 1, 2, 3, \ldots$, then the phase shift acquired by the charge going around this curve is $n\Delta\phi$, independent of the shape of this curve. (Topology has to do with integers. So, a topological phase shift should, strictly speaking, be expressed in terms of integers.) Then the curve $-\mathbf{r}(t)$, which is the parity transform of the original curve, goes around the line charge n times in the dual situation also. Therefore, the phase shift acquired in the dual situation is $n\Delta\phi$, independent of the shape of this curve. Hence, the latter phase shift is also topological. This may also be seen from the fact that the expressions (2.1) and (2.5) for these phase shifts are independent of the shape of the curve traveled by the particle. But notice that the argument above which establishes the equality of phase shifts in the two situations that are dual to each other does *not* assume that the phase shift is topological. It is valid for the phase shift due to any interaction, which may or may not be topological. Also, the above duality can be generalized to the case of the charge moving around a closed curve C and acquiring a phase in the field of an arbitrary distribution of dipoles, each having the same magnetic moment in both direction and magnitude. A little thought, by considering each pairwise interaction of the charge and each dipole, shows that in the dual situation, in which the dipole moves along the parity transformed curve with the charges in the parity transformed positions of the dipoles of the original situation, the same phase is acquired by the dipole. Again using the invariance of this interaction under parity, it is concluded that the same phase shift is obtained when the dipole travels the original closed curve C with the charges in the positions of the dipoles in the original situation [9]. This argument may be generalized to an arbitrary linear interaction, but the interaction needs to be invariant under parity for the last step to be valid. The equality between the two phases in the two dual situations will be called the duality principle.

This duality principle enables us to obtain from the known phase shift due to a line source a new phase shift. Alternatively, if both phase shifts are known then this principle may be used heuristically to obtain a new solution of the field equations from the old solution that gave the old phase shift. For example, suppose we know the magnetic field of a solenoid and the AB phase shift [1] of a charge due to it, and the phase shift of a magnetic moment due to a general electric field [2]. Then according to the duality principle, the phase shift in the situation dual to the AB effect in which a charge is interfering in the electric field $\mathbf{E}$ is due to a line charge is the same and is given by (2.5). Then using the result for the phase shift of the magnetic moment due to $\mathbf{E}$ implied by (1.4), and the axial symmetry, $\mathbf{E} = \frac{\lambda}{2\pi\rho}\hat{\rho}$, where ρ is the distance from the line charge, $\hat{\rho}$ is a unit vector in the radial direction and λ is the charge per unit length. Thus $\mathbf{E}$ of a line charge is obtained without solving Maxwell's equations. In the next section, I shall apply this general argument to phase shifts produced by gravity.

3. Topological Phase Shifts in the Gravitational Field

It is well known that the mass and the spin angular momentum in a gravitational field are, respectively, analogous to the charge and magnetic moment in an electromagnetic field. Therefore the analog of the AB effect for gravity is the phase shift $\Delta\phi_G$ acquired by a mass going around a string that has angular momentum (analog of the solenoid). The dual situation then is a spin going around a string or a rod having mass only, and acquiring a phase $\Delta\phi'_G$. Then according to the general argument in section 1, if the field equations are linear,

$$\Delta\phi'_G = \Delta\phi_G. \tag{3.1}$$

The actual values of $\Delta\phi_G$ and $\Delta\phi'_G$ depends on the gravitational theory used to compute them. I shall study these phase shifts in the following theories: A. Newtonian gravity, B. linearized limit of Einstein's theory of general relativity, and C. The Einstein-Cartan-Sciama-Kibble (ECSK) theory of the gravitational field with torsion [14]. In all three cases (3.1) will be shown to be satisfied. The differences between these phase shifts provide a way of distinguishing, in principle, between these theories, although in practice the predicted effects are too small for realistic experimental tests at the present time.

A. Newtonian Gravity

In this case, only the mass, not the angular momentum, acts as the source of gravity and is acted upon by gravity. Therefore, both $\Delta\phi_G$ and $\Delta\phi'_G$ are zero. Hence, (3.1) is trivially satisfied.

B. Linearized General Relativity

Consider now the low energy weak field limit of general relativity. Write the metric as $g_{\mu\nu} = \eta_{\mu\nu} + \gamma_{\mu\nu}$, where $\gamma_{\mu\nu} << 1$. In this subsection, all terms which are second order in $\gamma_{\mu\nu}$ will be neglected. On writing $\overline{\gamma}_{\mu\nu} = \gamma_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}\eta^{\alpha\beta}\gamma_{\alpha\beta}$, the well known linearized Einstein field equations are

$$\partial^{\alpha}\partial_{\alpha}\overline{\gamma}_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{3.2}$$

in the gauge defined by $\partial^{\nu}\overline{\gamma}_{\mu\nu} = 0$. I neglect stresses so that we have

$$T_{ij} = 0, \overline{\gamma}_{ij} = 0, i, j = 1, 2, 3. \tag{3.3}$$

Consider now a particle with mass m and intrinsic spin $\mathbf{S}$ at low energies. In the stationary situation, a coordinate system may be chosen so that $\gamma_{\mu\nu}$ are time independent. Then the acceleration due to gravity is $\mathbf{g} = -\frac{1}{2}\nabla\gamma_{00}$ and the 'Coriolis' vector potential is $\gamma_0 = -(\gamma_{01}, \gamma_{02}, \gamma_{03})$, as can be seen easily from the geodesic equation. The 'gravi- magnetic field' $\mathbf{H} = \nabla \times \gamma_0 = 2\mathbf{\Omega}$, where $\mathbf{\Omega}$ is the angular velocity of the coordinate basis relative to the local inertial frame. To couple the spin to the gravitational field introduce the vierbein $e^\mu{}_a$ and its inverse $e_\mu{}^a$:

$$e^\mu{}_a = \delta^\mu{}_a - \frac{1}{2}\gamma^\mu{}_a, e_\mu{}^a = \delta_\mu{}^a + \frac{1}{2}\gamma_\mu{}^a, \tag{3.4}$$

which satisfy (3.20) and (3.22) below. The latin indices, which take values $0, 1, 2, 3$ may now be lowered and raised using the Minkowski metric η_{ab} and its inverse η^{ab}. It then follow hat the Ricci rotation coefficients $\omega_\mu{}^a{}_b \equiv e_\nu{}^a \nabla_\mu e^\nu{}_b$ are given by

$$\omega_{\mu ab} = \frac{1}{2}(\gamma_{\mu b,a} - \gamma_{\mu a,b}), \tag{3.5}$$

where $,a$ denotes partial differentiation with respect to x^a.

The phase shift in interference due to the gravitational field may be obtained in the present approximation by taking the low energy weak field limit of the phase shift obtained in reference 5. In particular, the phase shift due to spin alone is obtained by parallel transporting the spin wave function by acting on it by the operator

$$\Phi_S = P \exp\left[-\frac{i}{\hbar}\int_C \frac{1}{2}\omega_\mu{}^a{}_b S^b{}_a dx^\mu\right] = P \exp\left[-\frac{i}{\hbar}\int_C \frac{1}{2}\gamma_{\mu a,b} S^{ab} dx^\mu\right], \tag{3.6}$$

where the integral is along the unperturbed classical trajectory with P denoting path ordering, and $S^b{}_a$, which generate Lorentz transformations in spin space, are related to the spin vector S^a and the 4- velocity v^b by

$$S^{ab} = \epsilon^{abcd} v_c S_d, \tag{3.7}$$

with all components being with respect to the vierbein. The subsidiary condition $S^a v_a = 0$ is assumed here.

Rewrite (3.6) as

$$\Phi_S = P \exp\left[-\frac{i}{\hbar}\oint_C \frac{1}{2}\gamma_{0a,b} S^{ab} dt - \frac{i}{\hbar}\oint_C \frac{1}{2}\gamma_{ia,b} S^{ab} dx^i\right]. \tag{3.6'}$$

The first integral in the exponent of (3.6′) is

$$\oint_C \frac{1}{2}\gamma_{00,i} S^{0i} dt + \oint_C \frac{1}{2}\gamma_{0i,j} S^{ij} dt = \oint_C (\mathbf{g} \times \mathbf{S} \cdot \mathbf{v} - \mathbf{\Omega} \cdot \mathbf{S}) dt. \tag{3.8}$$

The second integral of (3.6′) is approximately

$$\oint_C \frac{1}{2}\gamma_{ij,k} S^{jk} dx^i = \oint_C \mathbf{g} \times \mathbf{S} \cdot d\mathbf{r}. \tag{3.9}$$

Combining these results, (3.6) reads

$$\Phi_S = P\exp\left[-\frac{i}{\hbar}2\oint_C \mathbf{g}\times\mathbf{S}\cdot d\mathbf{r} + \frac{i}{\hbar}\oint_C \mathbf{\Omega}\cdot\mathrm{S}dt\right]. \tag{3.10}$$

The precession represented by the last term of the exponent in (3.10) corresponds to the interaction energy $-\mathbf{\Omega}\cdot\mathbf{S} = -\frac{1}{2}\mathbf{S}\cdot\mathbf{H}$ in the Hamiltonian. This may be understood from the fact that when we transform to a frame rotating with an angular velocity $\mathbf{\Omega}$ relative to the local inertial frame, the spin that is constant in the inertial frame obviously rotates with angular velocity $-\mathbf{\Omega}$ relative to the new frame [15]. The ratio γ of the magnetic moment to spin introduced in section 1 for the electromagnetic interaction is, for a particle with charge e and mass m, $\gamma = \frac{ge}{2m}$, where g is the gyromagnetic ratio. For the gravitational field, the principle of equivalence implies that the 'charge' density equals mass density. Therefore, $g = 1$ and $e = m$. Hence, $\gamma = \frac{1}{2}$ and the 'gravi- magnetic moment' $\mu_G = \frac{1}{2}\mathbf{S}$, consistent with the above interaction energy.

Equations (3.2) subject to (3.3) are

$$\partial^\alpha\partial_\alpha\bar{\gamma}_{0\mu} = 8\pi GT_{0\mu}.$$

These are like Maxwell's equations in the Lorentz gauge, and may be solved in the same way. Consider the specific case of an infinite uniform hollow cylinder of radius ρ_0 and mass per unit length μ rotating about its axis with angular momentum per unit length J parallel to the axis of the cylinder along the z-axis. This is analogous to a rotating charged cylinder in electromagnetism. So, on defining $\mathbf{r} = (x^1, x^2, x^3)$, $\rho = (x^1, x^2, 0)$, and $\rho = |\rho|$, the solution exterior to the cylinder ($\rho > \rho_0$) is obtained to be

$$\gamma_{00} = \gamma_{11} = \gamma_{22} = \gamma_{33} = 4G\mu\log\rho, \gamma_0 = -\frac{4G}{\rho^2}\mathbf{J}\times\mathbf{r} = -\frac{4G}{\rho}\mathbf{J}\times\hat{\rho}, \tag{3.11}$$

where $\hat{\rho}$ is a unit vector in the direction of ρ. The solution in the interior to the cylinder is

$$\gamma_{00} = \gamma_{11} = \gamma_{22} = \gamma_{33} = 4G\mu\log\rho_0, \gamma_0 = -\frac{4G}{\rho_0^2}\mathbf{J}\times\mathbf{r} = -\frac{4G}{\rho_0^2}\mathbf{J}\times\rho. \tag{3.12}$$

Suppose at first that $J = 0$. Then, from (3.11), $\mathbf{\Omega} = 0$ and $\mathbf{g} = -\frac{2G\mu}{\rho}\hat{\rho}$. Consider the interference around the cylinder of a particle whose spin is polarized in the x^3–direction with the axis of the cylinder lying along the x^3- axis. Then the phase shift due to the coupling of spin to curvature [5,16] is obtained from (3.10) to be

$$\Delta\phi_G = -\frac{2}{\hbar}\oint_C \mathbf{g}\times\mathbf{S}\cdot d\mathbf{r} = -8\pi\frac{G}{\hbar}\mu S. \tag{3.13}$$

This phase shift is independent of C and is therefore topological.

Consider now the dual situation that is constructed as follows. Divide the cylinder into small segments of length $\delta\ell$. The mass of each segment is $m = \mu\delta\ell$.

In performing the duality operation, each segment is replaced by a segment whose spin is the same as S and the particle is replaced with another particle with mass m. Then the cylinder has angular momentum per unit length $J = \frac{S}{\delta l}$. Therefore,

$$\frac{m}{\mu} = \frac{S}{J}. \tag{3.14}$$

So, in the dual situation, the mass m is interfering around the cylinder with angular momentum per unit length $\mathbf{J}$, which is a gravitational analog of the AB experiment. From (3.11), the phase shift due to $\mathbf{J}$ is the Sagnac phase shift [4]

$$\Delta\phi'_G = \frac{m}{\hbar}\oint_C \gamma_0 \cdot d\mathbf{r} = -8\pi\frac{G}{\hbar}mJ. \tag{3.15}$$

It follows from (3.13), (3.14) and (3.15) that (3.1) is satisfied.

I shall now describe the gravitational analog to the topological phase shift of the neutron due to the magnetic field described towards the end of section 1. Suppose, in a neutron or atomic interference experiment each of the two interfering beams passes along the axis of each of two identical very massive cylinders. One of the cylinders rotates as the wave packet of each neutron enters the cylinder and stops rotating before the neutron leaves the cylinder. Then from (3.12),

$$\mathbf{\Omega} = \frac{1}{2}\nabla \times \gamma_0 = -\frac{4G}{{\rho_0}^2}\mathbf{J}. \tag{3.16}$$

Suppose, for simplicity, that the spin of the neutron or atom is polarized along the axis of the cylinder. The first integral in (3.10) would be the same for both beams. Therefore, the phase shift between them is given by the second integral of (3.10) due to the rotating cylinder to be

$$\Delta\phi_\Omega = \frac{1}{\hbar}\oint_C \mathbf{\Omega}\cdot\mathbf{S}dt = -\frac{4GJS\tau}{\hbar{\rho_0}^2}, \tag{3.17}$$

on using (3.16), where τ is the time spent by the neutron inside the hollow cylinder, assuming that the time intervals during which the rotation of the cylinder is turned on and off is negligible compared to τ. As in the electromagnetic case, the phase shift (3.17) is not accompanied by a force, apart from transient effects when the field is turned on and off which occurs also for the AB effect due to the scalar potential [1]. Hence, (3.17) is a topological phase shift.

It is interesting that $\Delta\Phi_S$ obtained here by parallel transport with respect to the gravitational connection is analogous to how the electromagnetic phase shift experienced by the dipole was obtained by parallel transport with respect to a corresponding connection [8,9]. The three gravitational phase shifts obtained above using (3.10) and (3.15) are the low energy weak field limit of the phase shifts obtained previously using Dirac's equation [5]. However, these phase factors correspond to the tentative Hamiltonian

$$H = \frac{1}{2m}(\mathbf{p} - m\gamma_0 - 2\mathbf{S}\times\mathbf{g})^2 + m\frac{\gamma_{00}}{2} - \frac{1}{2}\mathbf{S}\cdot\mathbf{H}, \tag{3.18}$$

in the sense that they may be derived from this Hamiltonian. This is a generalization of the Hamiltonian found by DeWitt [17] to include spin. One way of confirming (3.18) is to directly take the low energy weak field limit of Dirac's equation, which will be studied in a future paper.

C. General Relativity with Torsion

The phase shift in general relativity may be obtained from the action on the wave function of the gravitational phase factor [18]

$$\Phi_C = P\exp[-i\int_C (e_\mu{}^a P_a + \frac{1}{2}\omega_\mu{}^a{}_b M^b{}_a)dx^\mu], \tag{3.19}$$

where C goes through the interfering beams. Here, P_a and $M^b{}_a, a, b = 0,1,2,3$ are the translation and Lorentz transformation generators which generate the Poincare group that acts on the Hilbert space. Then $e_\mu{}^a$ and $\omega_\mu{}^a{}_b$ have the interpretation of the gauge potentials of a Poincare gauge field. In an interferometry experiment the two beams need to be brought together by means of mirrors which gives rise to the Thomas precession [19], which will be treated elsewhere [20].

It was shown by means of the WKB approximation of Dirac's equation that (3.19) determines the gravitational phase [5,18]. This may also be realized for a particle with arbitrary spin as follows: The Lorentz part of (3.19) ensures that the wave packet is parallel transported infinitesimally, while it acquires a phase, which is a good approximation for the locally approximate plane wave being considered here. To find the phase acquired due to energy momentum, note first that $e_\mu{}^a$ depends on the observer. A Lorentz transformation of the observer results in $e_\mu{}^a$ transforming as a contravariant vector in the index a while P_a transforms as a covariant vector. Suppose that a particle is in a state $|\psi>$ that is approximately an eigenstate of P_a with eigenvalues p_a. The fact that the gravitational phase is observable along an open curve implies that the "wave vector" $p_\mu \equiv e_\mu{}^a p_a$ is observable [21,22]. Requiring that the correspondence between p_μ and p_a is $(1-1)$ implies that $e_\mu{}^a$ is a non singular matrix. Therefore, it has an inverse $e^\mu{}_b$:

$$e_\mu{}^a e^\mu{}_b = \delta^a{}_b. \tag{3.20}$$

Hence, $p_a = e^\mu{}_a p_\mu$. The Casimir operator $\eta^{ab}P_a P_b$ of the Poincare group has a definite value, say m^2, for the given particle. Therefore,

$$\eta^{ab}p_a p_b = g^{\mu\nu}p_\mu p_\nu = m^2, \tag{3.21}$$

where $g^{\mu\nu} \equiv \eta^{ab}e^\mu{}_a e^\nu{}_b$ is a non singular matrix. Its inverse $g_{\mu\nu}$ defines a pseudo-Riemannian metric of Lorentzian signature on space-time. On using (3.20),

$$g_{\mu\nu} = \eta_{ab}e_\mu{}^a e_\nu{}^b. \tag{3.22}$$

Thus the definiteness of the mass (which may be zero) ensures the definiteness of the phase that depends on $e_\mu{}^a$ even for open curves. In this way the space-time metric is deduced from the gravitational phase corresponding to the translational part of (3.19) which is observable along an open curve. Conversely, the metric determines the latter phase to be observable along an open curve.

The field strength or curvature of this Poincare gauge field is obtained by evaluating (3.19) for an infinitesimal closed curve C:

$$\Phi_C = 1 - \frac{i}{2}(Q_{\mu\nu}{}^a P_a + \frac{1}{2}R_{\mu\nu}{}^a{}_b M^b{}_a)d\sigma^{\mu\nu}, \tag{3.23}$$

using the Poincare Lie algebra, where

$$Q^a = de^a + \omega^a{}_b \wedge e^b \tag{3.24}$$

is the torsion and

$$R^a{}_b = d\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b \tag{3.25}$$

is the linear curvature.

Comparison of (3.19) with (1.2) and (1.4), and (3.23) suggest that there may be topological phase shifts due to interference of coherent beams that enclose a region that contains curvature and torsion, but which are zero along the beams. Such an example is provided by the cosmic string whose metric exterior to the string is given cylindrical coordinates as [23, 24, 25].

$$ds^2 = (dt + \beta d\phi)^2 - d\rho^2 - \alpha^2\rho^2 d\phi^2 - dz^2, \tag{3.26}$$

where α and β constants. Then the metric $g_{\mu\nu}$ satisfies (3.22) for the following orthonormal co-frame field $\{e^a\}$ adapted to the above coordinate system:

$$e^0 = dt + \beta d\phi, e^1 = d\rho, e^2 = \alpha\rho d\phi, e^3 = dz. \tag{3.27}$$

The connection coefficients in this basis are $\omega_\mu{}^a{}_b \equiv e_\nu{}^a \nabla_\mu e^\nu{}_b = 0$, for all a, b, μ except for

$$\omega_\phi{}^1{}_2 = -\omega_\phi{}^2{}_1 = -\alpha d\phi. \qquad (3.28))$$

It follows, on using (3.24) and (3.25), that $Q^a = 0, R^a{}_b = 0$ outside the string. The scattering cross section of particles with definite energy in the above geometry has been obtained before [26].

In the appendix it is shown that this solution may be extended to an interior solution that has uniform energy and spin densities and which generate curvature and torsion according to the ECSK equations [14]. The constants α and β are then determined by matching the interior and exterior solutions to be

$$\alpha = 1 - 4G\mu, \beta = 4GJ, \tag{3.29}$$

where μ is the mass per unit length and J is the angular momentum per unit length due to the intrinsic spin density inside the string.

If C is a closed curve around the cosmic string then from (3.19),

$$\Phi_C = \exp\left(-i \oint_C e_\mu{}^0 P_0 dx^\mu\right) P \exp\left[-i \oint_C \left(\sum_{k=1}^{3} e_\mu{}^k P_k + \omega_\mu{}^1{}_2 M^2{}_1\right) dx^\mu\right], \qquad (3.30)$$

using the fact that $e_\mu{}^0 P_0$ commutes with the other terms that occur in (3.30).The first exponential is a time translation second is a spatial Euclidean transformation. Hence, if (3.23) is valid, then the time translation would correspond to torsion being non zero inside the string. Suppose that the surface of the string is given by $\rho = \rho_0$, where ρ_0 is a small constant. Then, substituting (3.27), (3.28) into (3.30), and using (3.23), the flux of torsion through a cross-section Σ of the string is

$$\int_\Sigma Q^0 = 2\pi\beta, \int_\Sigma R^1{}_2 = 2\pi(1-\alpha) \qquad (3.31)$$

This is independent of the particular geometry inside the string so long as Σ is "infinitesimal" so that (3.23) is valid. In particular, (3.31) is easily verified for the solution in the appendix, independently of the value of ρ_-, using (A.12), (A.15) and (A.16).

For simplicity, consider a circular interferometer with constant radius $r > \rho_0$ in a plane normal to the string, with its center on the axis of the string. It may be a superconducting interferometer, e. g. a superconducting ring interrupted by a Josephson junction. Or it may be an electron interferometer, or a wave guide, such as an optical fiber, at one point of which is the beam splitter that splits a beam into two which travel in opposite senses and interfere at a mirror that is at another point in the interferometer. The interferometer does not rotate relative to the distant stars, which may be ensured by requiring that telescopes rigidly attached to this interferometer are focused on the distant stars.

The phase shift may be obtained using (3.30) with C along integral curves of p^μ which lie on a 2 dimensional submanifold σ with constant z and $\rho = r$. Hence, C may be chosen to be along a circle around σ with constant t. Suppose E is the energy of the wave function which is assumed to be constant in time at the beam splitter. Then it is constant everywhere along the beams. Therefore, in this WKB approximation, the magnitude of the momentum $p = (E^2 - m^2)^{1/2}$ is also a constant along the beam. By taking into account the Fermi-Walker transport of vectors associated with $|\psi>$, $M^2{}_1$ in (3.30) may be replaced by the spin operator $S^2{}_1$ in the present coordinate basis [19]. The spin is assumed to be polarized in the z-direction, i. e. $|\psi>$ is an eigenvector of $S^2{}_1$ with eigenvalue $S/\hbar$.

Now, the three operators in (3.30) commute with one another and their actions on $|\psi>$ give rise to the following topological phase shifts: (i) The general relativistic Sagnac phase shift [4] is obtained from the first factor to be

$$\Delta\phi_E = -\oint_C \frac{E}{\hbar} e_{\hat{\phi}}{}^0 d\phi = -\int_\Sigma \frac{E}{\hbar} Q_{\hat{\rho}\hat{\phi}}{}^0 d\rho \wedge d\phi = -8\pi \frac{G}{\hbar} EJ, \qquad (3.32)$$

where Σ is a 2-surface spanned by C. (ii) The phase shift due to the coupling of spin to curvature [5] which is obtained by the second factor in (3.30)*

$$\Delta\phi_S = -\frac{S}{\hbar}\{\oint_C \omega_{\hat{\phi}}{}^1{}_2 d\phi - 2\pi) = -\int_\Sigma R_{\hat{\rho}\hat{\phi}}{}^1{}_2 d\rho \wedge d\phi = -8\pi\frac{G}{\hbar}S\mu. \tag{3.33}$$

The phase shifts (3.32) and (3.33) are expressed as flux integrals of torsion and curvature because they could have also been obtained from (3.19) which depends only on the affine connection. (The torsion and curvature fluxes contained in (3.32) and (3.33) are the same as (3.31) which is independent of the particular geometry interior to the string onsidered here.) It follows that these phase shifts are independent of the shape of the interferometer enclosing the string and therefore may be called topological.

I shall now show that the above topological effects satisfy the principle of duality formulated in section 2. Consider first the Sagnac effect on a particle with energy E due to the spinning string with angular momentum per unit length J. This is like the AB effect due to a solenoid. Divide the string into small segments of length $\delta\ell$. The spin of each segment is $S = J\delta\ell$. In performing the duality operation, each segment is replaced by a segment whose mass is the same as E and the particle is replaced with another particle with spin S. Then the solenoid has been replaced by a rod with mass per unit length $\mu = \frac{E}{\delta\ell}$. Therefore,

$$\frac{E}{\mu} = \frac{S}{J}. \tag{3.34}$$

Conversely, if (3.34) is valid then the two situations may be obtained from each other by performing the duality operation. Hence, by the duality principle, the phase shifts for the two situations should be equal. Indeed, the phase shifts (3.32) and (3.33) which were derived without paying any attention to the duality principle are equal if and only if (3.34) is valid.

This illustrates also again how the duality principle may be used to obtain the phase shift for the dual situation: From (3.32), we may obtain (3.33), or vice versa, on using (3.34). Even though the general relativistic equations are in general non linear, the equations that are solved in the appendix to obtain the exact solution are all linear, so that there must be duality in the present case according to the general arguments of section 2. If this duality is assumed then a new gravitational solution may be obtained from an old solution both in the present case and in the low energy weak field case considered earlier, similar to how this was done in the

* The phase shifts (3.32) and (3.33) may be evaluated using the line integral outside the string using (A.16) or the surface integral inside the string using (A.11). In (3.33), 2π has been subtracted from the line integral to remove the purely coordinate effect due to the rotation of e^2 by 2π as one goes around C, consistent with the Gauss-Bonnet theorem.

electromagnetic case at the end of section 2.

4. Concluding Remarks

As already mentioned, the AB effect shows that the field strength is insufficient to describe the electromagnetic field, whereas the phase factor (1.2), which is called a holonomy transformation because it parallel transports around a closed curve, adequately describes the field. More generally, for an arbitrary gauge field, the holonomy transformations are of the form (1.4), with $A_\mu{}^k$ now being the corresponding vector potential. The fact that these are sufficient to determine the field uniquely is shown by the following theorem [27]: Given the holonomy transformations (2) for piece-wise differentiable curves which begin and end at a given point in space-time, the gauge potential $A_\mu{}^k$ may be reconstructed, and it is then unique up to gauge transformations.

But since the set of such curves form an infinite dimensional manifold L, the corresponding operators (2) have a great deal of redundancy. Indeed, the gauge field in space-time may be reconstructed from a minimal set of these operators defined on a four dimensional submanifold of L [7]. This is mathematically equivalent to working with the gauge potential defined in a particular gauge on the four dimensional space-time. Therefore, once the redundancy in the loop space L has been removed, there is no advantage to using the holonomy transformation (2) as opposed to the gauge potential in a particular gauge.

It follows that in quantizing the electromagnetic or more general gauge fields, one must quantize the gauge potential instead of the field strength. Similarly, the topological effects due to the gravitational field described above suggest that in quantizing the gravitational field, it is the 'gauge potentials' $e_\mu{}^a$ and $M^b{}_a$ which should be quantized, and the metric (3.22) is obtained from them as a secondary variable [26,22]. However, there is a breaking of gauge symmetry which makes $e_\mu{}^a$ a tensor field instead of a connection [22]. This is like how in a superconductor the $U(1)$ gauge symmetry is spontaneously broken, which makes A_μ a covariant vector field instead of a connection.

So, it may well be that in the early universe there was the full Poincare gauge symmetry with $e_\mu{}^a$ and $M^b{}_a$ having vacuum expectation value zero in an appropriate gauge. As a result of spontaneous symmetry breaking of the translational part of the Poincare group, $e_\mu{}^a$ may have acquired a vacuum expectation value equal to $\delta_\mu{}^a$ corresponding to the Minkowski geometry. But I emphasize that these are speculative remarks, and need justification by a detailed theory.

Acknowledgements

I thank Y. Aharonov and P. O. Mazur for stimulating discussions and J. L. Safko for several enlightening discussions. This work was partially supported by

NSF grant PHY-9307708 and ONR grant R&T 3124141.

Appendix: Spinning Torsion String

The simplest gravitational field equations in the presence of torsion are the Einstein-Cartan-Sciama- Kibble (ECSK) equations [14], which may be written in the form [29]

$$\frac{1}{2}\eta_{ijkl}\theta^l \wedge R^{jk} = -8\pi G t_i, \tag{A.1}$$

$$\eta_{ijkl}\theta^l \wedge Q^k = 8\pi G s_{ij}, \tag{A.2}$$

where t_i and s_{ij} are 3-form fields representing the energy-momentum and spin densities. I shall now obtain an exact solution of these equations for the interior of the cosmic string which matches the exterior solution (3.26). This will then give physical and geometrical meaning to the parameters α and β in (3.26). This solution will be different from earlier torsion string solutions [31] that have static interior metrics matched with exterior metrics which are different from (3.26).

The ρ and z coordinates in the interior will be chosen to be the distances measured by the metric in these directions. Since the exterior solution has symmetries in the t, ϕ, and z directions, it is reasonable to suppose the same for the interior solution. So, all functions in the interior will be functions of ρ only. Requiring also simplicity, I make the following ansatz in the interior:

$$\theta^0 = u(\rho)dt + v(\rho)d\phi, \theta^1 = d\rho, \theta^2 = f(\rho)d\phi, \theta^3 = dz, \omega^2{}_1 = k(\rho)d\phi = -\omega^1{}_2, \tag{A.3}$$

all other components of $\omega^a{}_b$ being zero, and $ds^2 = \eta_{ab}\theta^a\theta^b \equiv g_{\mu\nu}dx^\mu dx^\nu$. Suppose also that the energy density ϵ and spin density σ, polarized in the z-direction, are constant and correspond to a classical fluid at rest. I. e.

$$\begin{aligned} t_0 &= \epsilon\theta^1 \wedge \theta^2 \wedge \theta^3 = \epsilon f(\rho)d\rho \wedge d\phi \wedge dz, \\ s_{12} = -s_{21} &= \sigma\theta^1 \wedge \theta^2 \wedge \theta^3 = \sigma f(\rho)d\rho \wedge d\phi \wedge dz, \end{aligned} \tag{A.4}$$

the other components of s_{ij} being zero. In terms of the components of the energy-momentum and spin tensors in the present basis, this means that $t^0{}_0 = \epsilon =$ constant and $s^0{}_{12} = \sigma =$ constant.

It is assumed that there is no surface energy-momentum or spin for the string. Then the metric must satisfy the junction conditions [30], which in the present case are

$$g_{\mu\nu}|_- = g_{\mu\nu}|_+, \partial_{\hat\rho} g_{\mu\nu}|_+ = \partial_{\hat\rho} g_{\mu\nu}|_- + 2K_{(\mu\nu)\hat\rho}, \tag{A.5}$$

where $K_{\alpha\beta\gamma} = \frac{1}{2}(-Q_{\alpha\beta\gamma} + Q_{\beta\gamma\alpha} - Q_{\gamma\alpha\beta})$ is the contorsion or the defect tensor, $|_+$ and $|_-$ refer to the limiting values as the boundary of the string is approached from outside and inside the string,, respectively, and the hat denotes the corresponding coordinate component.

Substitute (A.3), (A.4) into the Cartan equations (A.2). The $(i,j) = (0,2)$, $(0,3),(2,3)$ eqs. are automatically satisfied. The $(i,j) = (0,1),(1,3),(1,2)$ eqs. yield

$$f'(\rho) - k(\rho) = 0, u'(\rho) = 0, v'(\rho) = 8\pi G\sigma f(\rho), \tag{A.6}$$

where the prime denotes differentiation with respect to ρ. Therefore, the continuity of the metric (eq. (A.5)) implies that, since $u = 1$ at the boundary, $u(\rho) = 1$ everywhere. Now substitute (A.3), (A.4) into the Einstein equations (A.1). The $i = 0$ eq. yields

$$k'(\rho) = -8\pi G\epsilon f(\rho). \tag{A.7}$$

The $i = 1,2,3$ equations yield, respectively

$$t_1 = 0, t_2 = 0. t_3 = \frac{k'}{8\pi G} dt \wedge d\rho \wedge d\phi = -\epsilon\theta^0 \wedge \theta^1 \wedge \theta^2, \tag{A.8}$$

using (A.7). Hence, $t^3{}_3 = \epsilon = t^0{}_0$. From (A.6) and (A.7),

$$f''(\rho) + \frac{1}{\rho*^2} f(\rho) = 0, \tag{A.9}$$

where $\rho* = (8\pi G\epsilon)^{-1/2}$. In order for there not to be a metrical "cone" singularity at $\rho = 0$, it is necessary that $\theta^2 \sim \rho d\phi$ near $\rho = 0$. Hence, the solution of (A.9) is $f(\rho) = \rho * sin \frac{\rho}{\rho*}$. Then from (A.6), $k(\rho) = cos \frac{\rho}{\rho*}$, and requiring $v(0) = 0$ to avoid a conical singularity, $v(\rho) = 8\pi G\sigma\rho *^2 \left(1 - cos \frac{\rho}{\rho*}\right)$. This gives the metric in the interior of the string to be

$$ds^2 = \left[dt + 8\pi G\sigma\rho *^2 \left(1 - cos \frac{\rho}{\rho*}\right)\right]^2 - d\rho^2 - \rho *^2 \, sin^2 \left(\frac{\rho}{\rho*}\right) d\phi^2 - dz^2. \tag{A.10}$$

The only non vanishing components of curvature and torsion are

$$Q^0 = 8\pi G\sigma\rho * \, sin \left(\frac{\rho}{\rho*}\right) d\rho \wedge d\phi, R^1{}_2 = \frac{1}{\rho*} sin \left(\frac{\rho}{\rho*}\right) d\rho \wedge d\phi = -R^2{}_1. \tag{A.11}$$

I apply now the junction conditions (A.5), which will show that ρ is discontinuous across the boundary. Denote the values of ρ for the boundary in the internal and external coordinate systems by ρ_- and ρ_+ respectively. ¿From (3.22) and (A.10), $g_{\hat{t}\hat{\phi}}$ and $g_{\hat{\phi}\hat{\phi}}$ are respectively continuous iff

$$\beta = 8\pi G\sigma\rho *^2 \left(1 - cos \frac{\rho_-}{\rho*}\right), \tag{A.12}$$

$$\alpha\rho_+ = \rho * \, sin \frac{\rho_-}{\rho*}. \tag{A.13}$$

The remaining metric coefficients are clearly continuous. The only non zero contorsion terms which enter into (A.5) are obtained from (A.11) to be

$$K_{(\hat{\phi}\hat{t})\hat{\rho}} = -4\pi G\sigma\rho * \, sin \frac{\rho}{\rho*}, K_{\hat{\phi}\hat{\phi}\hat{\rho}} = -(8\pi G\sigma)^2 \rho *^3 \left(1 - cos \frac{\rho}{\rho*}\right) sin \frac{\rho}{\rho*}. \tag{A.14}$$

Using (A.13) and(A.14), it can now be verified that the remaining junction conditions (A.5) are satisfied provided $\alpha = cos\frac{\rho_-}{\rho*}$. The mass per unit length is

$$\mu \equiv \int_\Sigma \epsilon\theta^1 \wedge \theta^2 = \frac{1}{4G}\left(1 - cos\frac{\rho_-}{\rho*}\right) = \frac{1}{8\pi G}\int_\Sigma R^1{}_2, \qquad (A.15)$$

where Σ is a cross-section of the string (constant t, z). Therefore, $\alpha = 1 - 4G\mu$. The angular momentum per unit length due to the spin density is

$$J \equiv \int_\Sigma \sigma\theta^1 \wedge \theta^2 = 2\pi\sigma\rho *^2\left(1 - cos\frac{\rho_-}{\rho*}\right) = \frac{1}{8\pi G}\int_\Sigma Q^0. \qquad (A.16)$$

Hence, from (A.12), $\beta = 4GJ$. The Sagnac phase shift obtained earlier is therefore $\Delta\phi = ET$, where T is the flux of Q^0 through Σ. In the special case when torsion is absent, which in the ECSK theory means that spin density is zero, $\beta = 0$, and the above solution reduces to the exact static solution of Einstein's theory found by Gott [32] and others [33], whose linearized limit was previously found by Vilenkin [34]. After this work was completed I learned that Tod [35] has studied torsion singularities using affine holonomy and the ECSK equations analogous to the present approach.

Note added in proofs: The phase shift due to spin in the interference around a rod and a cosmic string has also been studied by B. Reznik (PhD thesis, Tel Aviv Univ., 1994, and preprint to be published in Phys. Rev. D), by using the contribution to the Lagrangian due to the gravitational interaction energy $U = \frac{1}{2}\int T^{\mu\nu}\gamma_{\mu\nu}d^3x$. This amounts to treating gravity as a spin 2 field, compared to the present geometric approach which begins with the full general relativistic theory. However, the above mentioned paper assumes that T^{0i} in the rest frame of the particle is the curl of spin density, which is then boosted to the laboratory frame. This assumption corresponds to setting the 'gravi- magnetic moment' μ_G equal to the spin. This differs from the result in section 3 of the present paper that μ_G is half the spin in accordance with the principle of equivalence. The latter result implies that T^{0i} in the rest frame is half the curl of spin density. Then, integrating by parts, it is easy to show that the Lagrangian for a particle with mass m, velocity $\mathbf{v}$ and spin $\mathbf{S}$ in the laboratory frame is

$$L \equiv \frac{1}{2}m\mathbf{v}^2 - U = \frac{1}{2}m\mathbf{v}^2 - m\frac{\gamma_{00}}{2} + m\mathbf{v}\cdot\gamma_0 + 2\mathbf{v}\cdot\mathbf{S}\times\mathbf{g} + \frac{1}{2}\mathbf{S}\cdot\mathbf{H}$$

This confirms the Hamiltonian (3.18) of the present paper. Also, the present paper studies an additional spin interaction represented by the last term of (3.18). And this gives rise to the new topological phase shift (3.17).

References

1. Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959).

2. J. Anandan, Phys. Rev. Lett. **24**, 1660 (1982).
3. Y. Aharonov and A. Casher, Phys. Rev. Lett. **53**, 319 (1984).
4. J. Anandan, Phys. Rev. D **15**, 1448 (1977).
5. J. Anandan, Nuov. Cim. A **53**, 221 (1979).
6. T. T. Wu and C. N. Yang, Phys. Rev. D, **12**, 3845 (1975).
7. J. Anandan, Phys. Rev. D **33**, 2280 (1986).
8. J. Anandan, Phys. Lett. A **138**, 347 (1989).
9. J. Anandan in *Proceedings of the International symposium on the Foundations of Quantum Mechanics*, Tokyo, August 1989, edited by S. Kobayashi et al (Physical Society of Japan, 1990) P. 98.
10. A. S. Goldhaber, Phys. Rev. Lett. **62**, 380 (1989).
11. D. Wisnievsky and Y. Aharonov, Ann. Phys. **45**, 479 (1967).
12. A. Zeilinger in *Fundamental Aspects of Quantum Theroy*, edited by V. Gorini and A. Frigerio (Plenum, NY 1985).
13. B. E. Allman et al, Phys. Rev. Lett. **68**, 2409 (1992); errata, Phys. Rev. Lett. **70**, 250 (1993).
14. D. W. S. Sciama in Recent Developments in General Relativity (Oxford 1962). p. 415; T. W. B. Kibble, J. Math. Phys. **2**, 212 (1961).
15. See, for example, J. Anandan, Phys. Rev. Lett. **68**, 3809 (1992).
16. J. Anandan and B. Lesche, Lettre al Nuovo Cimento **37**, 391 (1983).
17. B. DeWitt, Phys. Rev. Lett. **16**, 1092 (1966).
18. J. Anandan in *Quantum Theory and Gravitation*, edited by A. R. Marlow (Academic Press, New York 1980), p. 157.
19. J. Anandan, Phys. Rev. D **24**, 338 (1981) sect. 3.
20. J. Anandan, to be published in Phys Lett. A (1994).
21. J. Anandan in *Topological Properties and Global Structure of Space-Time*, eds. P. G. Bergmann and V. De Sabbata (Plenum Press, NY 1985), p. 1-14; J. Anandan, Ann. Inst. Henri Poincare **49**, 271 (1988).
22. J. Anandan in Directions in Directions in General Relativity, Volume 1, Papers in honor of Charles Misner, edited by B. L. Hu, M. P. Ryan and C. V. Vishveshwara (Cambridge Univ. Press, 1993).
23. L. Marder, Proc. Roy. Soc. A **252**, 45 (1959) and in Recent Devolopments in General Relativity (Pergamon, New York 1962); A. Staruskiewicz, Acta. Phys. Pol. **424**, 734 (1963); J. S. Dowker, Nuov. Cim. B **52**, 129 (1967); J. L. Safko and L. Witten, Phys. Rev. D **5**, 293 (1972); V. B. Bezerra, J. Math. Phys. **30**, 2895 (1989) .
24. S. Deser, R. Jackiw, and G. 't Hooft, Ann. Phys. NY **152**, 220 (1984);
25. P. O. Mazur, Phys. Rev. Lett. **57**, 929 (1986).
26. P. O. Mazur, Phs. Rev Lett. **59**, 2380 (1987).
27. J. Anandan in *Conference on Differential Geometric Methods in Physics*, edited by G. Denardo and H. D. Doebner (World Scientific, Singapore, 1983) p. 211.

28. J. Anandan, Found. Phys. **10**, 601 (1980).
29. A. Trautman in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Holland, 1973).
30. W. Arkuszewski, W. Kopczynski, and V. N. Ponomariev, Commun. Math. Phys. **45**, 183 (1975).
31. A. R. Prasanna, Phys. Rev. D **11**, 2083 (1975); D. Tsoubelis, Phys. Rev. Lett. **51**, 2235 (1983).
32. J. R. Gott III, Astrophys. J. **288**, 422 (1985).
33. W. A. Hiscock, Phys. Rev. D **31**, 3288 (1985); B. Linet, Gen. Rel. and Grav. **17**, 1109 (1985).
34. A. Vilenkin, Phys. Rev. D **23**, 852 (1981).
35. K. P. Tod, Class. and Quantum Grav. **11**, 1331 (1994).

Disordered Gravitation: Localization and Diffusion Limited Dynamics of the Early Universe

Timir Datta

and

John L. Safko

Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208, USA

Abstract

Many quantum and classical fields are known to be influenced by disorder. Anderson transition of the Schrodinger field is a remarkable example of such a disorder effect. In this article we will discuss the effects of disorder on gravitation. The general relativistic (GR) gravitational field is a specially important case; because, this problem has not yet received much attention and disorder has to be introduced in a frame independent geometric manner. Furthermore, the GR equations are non-linear. Since gravity itself acts as a source of gravity, non-linear self-coupling is expected to cause the gravitational field to be more sensitive to disorder. Hence, any effects due to a random source distribution will amplify or "pile up" and encourage localization.

In this paper, we propose a simple model of a gravitating system with random disorder. Disorder is introduced through a stochastic generating function. The affine parameter, τ, for the out going null geodesics is calculated and observed to lengthen with the increase in disorder. We interpret this as a slowing down in the propagation of gravitational field and (as in the cases of other disordered fields) due to localization. We argue that localization must have been important in the early universe; when, due to thermal and quantum fluctuations, space time was very strongly disordered. In such a GR-localized era the cosmological scale factor, S(t), will be diffusive and slow. Our calculations show, in general, $S \sim t^{1/\gamma}$ where $\gamma > 2$. This random walk-like field propagation will effectively increase the value of the Newtonian constant G that can render the gravitational interaction to be strong enough to produce the nucleation sites for primordial matter. We reason, such random condensation could have been responsible for the observed inhomogeneity of the matter distribution in the present day universe.

1. Introduction

A system is said to be ordered when it has some symmetry; random breaking of symmetry causes the system to be disordered. Many new phenomena are known to result from disorder. The problem of random transitional symmetry breaking in a one dimensional phonon field was first reported by Dyson[1]. About a decade later Anderson considered a disordered electronic system[2,3]. Since that time, particularly after Anderson, many researchers have considered the case of disorder in both classical and quantum fields[4-7].

In this article we will consider the general behavior of all types of disorder induced localization phenomena with the goal to understand the peculiarities of the gravitational field localization. In particular, in the following sections we will briefly review the classic work by Anderson on the localization of electron waves. Later, the physical process of back and multiple scatterings in a strongly disordered system will be described. Finally the results from our model for gravitation with random disorder and the dynamics of a "GR-localized" universe will be discussed.

2. The Anderson Problem

Anderson studied the non-relativistic Schrodinger field of an electron in a random potential (disordered crystal) lattice. He reasoned that, in the absence of perfect translational symmetry the electronic eigenstates would not follow the Block-theorem. Hence, the wave vector **k** would cease to be a good quantum number and the wave function Φ would not belong to a unique wave vector. Further, for strong enough disorder the wave may be localized in space. Anderson determined that the envelope of such a localized state is peaked about its localization center and decreases rapidly away from that point. The probability of finding a localized electron rapidly approaches zero, as a function of the distance from its center.

Quantitatively, the problem consisted of two parts: the introduction of disorder in a mathematically tractable fashion while retaining the essential physics and the definition of a calculable parameter that measured the effect of disorder on the field. The first part was answered by modeling the unperturbed medium as a perfectly spaced lattice of uniform square well potential field V(x) of depth equal to V_0. On this lattice, a random potential W(x) was superposed. The width of the random distribution was W, as shown in Fig. 1a. For the second part Anderson proposed using the value of D(W,R), the quantum diffusion coefficient. D(W,R) is a measure of localization; that is, the absence of diffusion to an infinitely distant point is the criterion of complete localization. He showed, for sufficiently strong disorder, i.e., W/V_0 bigger than a critical value Δ^*, (the exact value of Δ^* being geometry and model dependent) $D(W,R) \rightarrow 0$ as $R \rightarrow \infty$.

This is behavior is known as the Anderson transition. At this transition, an electronic system undergoes a rapid change from some states localized to all states localized. This is a cooperative effect brought about by the coherent interference from all parts of the system. In practice, interference from the local regions is dominant; that is, constructive scatterings from the local sites are conducive to extended states. On the other hand, locally destructive interference or back-scatterings cause localization.[4] In the Anderson phase, disorder is strong and electron transport is absent. Figure 1b shows the extended, non-Block type electronic wave functions and Fig. 1c shows a localized wave function.

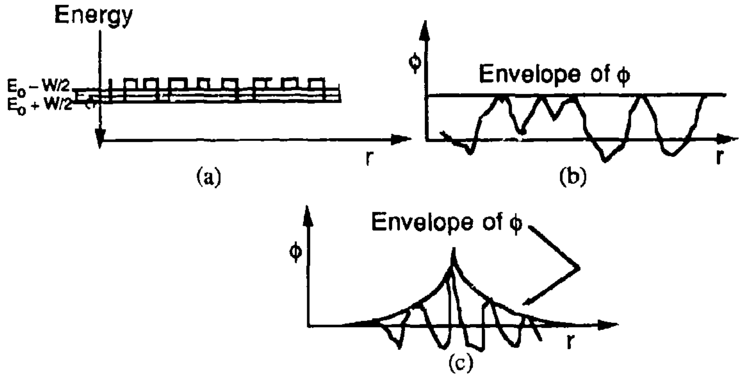


Figure 1. (a) Periodic Potential E_0 with applied random potential W.
(b) Non-Block type extended electron wave function.
(c) Localized wave.

Note that, because of coulomb attraction between the electron and the positive ion cores, the electron is trivially expected to be attached (localized) to the lattice points. The first unexpected quantum mechanical result was that, despite the scatterings from the zero-point vibrations of the lattice, the electronic states are extended. Furthermore, this behavior is independent of the value of the lattice constant. In the early years of the twentieth century, the effects of scattering due to zero-point vibration was a much contested question. The pursuit for evidence of zero-point scattering was a motivation for many X-ray diffraction experiments in crystals and low temperature resistance studies in pure metals. As lower and lower temperatures were achieved the research on resistance eventually led to the unexpected and startling discovery of superconductivity.

The question of lattice spacing was answered much later by Mott. Starting with the metallic state of extended wave eigenfunctions, Mott showed that as the lattice constant is increased the electron will minimize its energy by forming a bound state with a positive core. This will collapse the wave function into a hydrogenic orbital and result in a metal-insulator transition. Confinement costs kinetic energy but in the low density (large lattice spacing) limit, potential energy becomes the determining factor for the ground state configuration.

Another related transition was discussed by Wigner. The Wigner crystallization (localization) also takes place in the low density limit but does not require a lattice. This behavior is possible even in a uniform positive background of a "jellium" and is also primarily due to the coulomb correlation. There is some influence of exchange or Pauli-principal which tends to keep the electrons away from each other. Although, all three transitions relate to the change from bound to extended states, in the Mott and Wigner cases the underlying cause is the (classical) coulomb interaction. The remarkable aspect of the Anderson phenomenon is that it is due to (the translation) symmetry breaking.

The understanding of the electronic properties of disordered systems have resulted in many advances.[4-6] In the last ten years or so, these considerations have been extended to many quantum and non-quantum fields. For instance, localization has also been reported for the classical Maxwell (EM) field.[7]

It has been observed that, in suitably prepared disordered dielectric media and over some restricted frequency range, electromagnetic radiation can be localized. For such a medium, many transport properties such as microwave propagation has been shown to become slow and diffusive.[7] However, the localization criterion for vector waves such as the EM field is dependent on the details of the particular situation. Another reason for the difference between the electrons and photons is that the electron can be bound to a lattice site; however, the dielectric function of a medium has to be positive and real at all points, so a photon cannot form a locally bound state. In this context, because, of the universal attractive nature of gravitation, the GR problem is at the opposite limit from the photon case.

3. Physical Description of Localization

Even in a uniform material medium light propagates with a speed less than its speed in vacuum. So what is so special about localization? To make the distinction clear let us discuss the EM problem in some details.

Any medium has a substructure which makes it different from vacuum. This structure may be due to atoms or "grains" which affect (scatter) the incident field. In a "weak" medium the density (ρ) of these grains is small From one scatterer to the next the field propagates almost freely with a speed equal to c, just as in vacuum. However, by Huygens principal all the fields produced by all the scatters superposes and gives rise to a net field **F**(**R**,t). In absence of absorption the amplitude of **F** at position **R** remains the same as it would be in vacuum but **F**(**R**,t) acquires a change of phase, $\delta\phi$, proportional to the displacement **R.** This $\delta\phi$ is in addition to the phase increase due to the "distance effect" contribution in vacuum.[8] Hence, the effective wavelength ,λ, (the distance the field would require to travel for the total phase to undergo a full cycle or $\Lambda\phi = 2\pi$) is shorter in the medium than λ_0, the wave length in vacuum.

Clearly, there are several length scales of the medium. In the weakly interacting limit only three are important: R(t), the radius of the wave front (spherical in an assumed isotropic medium); the mean free path ℓ; and λ. In this limit, R >> ℓ >> λ. For scales much larger than λ and ℓ the medium may be treated as a continuum and the net effect is described by the refractive index **n**; where $\mathbf{n} = \lambda_0/\lambda$. The wave propagates uniformly at a constant rate. In a time dt, the displacement, **dR**, is given by the linear relation:

$$\mathbf{dR} = c^* \times dt \qquad [3.1]$$

where, $c^* = \lambda\nu = c/\mathbf{n}$, ν is the frequency, and $c^* < c$. This is the description of slow but constant rate of propagation in a uniform medium.

To understand localization, the coarse graining of the effective medium model described above needs to be relaxed. The problem has to be treated with a finer mesh by including two additional lengths, $< r >$ and ξ. The average intergrain distance, $< r > \sim (\rho^{1/d})$ where d is the Euclidean space-dimension and ρ is the density. The grains form clusters in the medium. The correlation length, ξ, is a measure of the typical cluster size. Also, ξ represents the scale length of disorder in the medium. At points separated by distances larger than ξ the medium is uncorrelated. The length ℓ is inversely dependent on the scattering cross section (Σ) and ρ. Hence, it represents the scale of interaction between the field and the medium. In the uniform effective medium limit described above $< r >$ and ξ do not play any role because , $\xi << < r > << \ell$.

Disorder effects are strongest and localization is possible if ξ, λ and ℓ are all of the same order but $\lambda \sim \ell < \xi$. With localization, the nature of wave propagation is qualitatively different! In this case ℓ is small and the field propagates in a random walk. Under this condition, the net distance traversed by the wave per unit time becomes progressively slow. Propagation slows down as the wave moves away from the source. In other words, the term "speed or rate of propagation" becomes ill-defined as **dR** becomes a sublinear function of dt, i.e.,

$$\mathbf{dR} = c^*(dt,dR) \times dt\,. \qquad [3.2]$$

The "speed" or the factor c* decreases with R; because, in each region of radius R, there are $N \sim \rho \times (R^d)$ scatterers and the number of scattering increases with R. A large number of these scatterings are strong enough to send the wave back into a region already traversed. Many back and forth, zigzag, random walk like steps are required to make a net forward displacement of the wavefront. Under complete localization, even after an infinitely long time the wave fails to arrive at infinity.

4. The Gravitational Problem

One may ask, why is the disorder problem of any interest in gravitation? There are two reasons: first, gravitation is one of the most pervasive interactions in nature and is known to have wavelike solutions, so it is reasonable to investigate gravitational localization. Second, at some early epoch in the history of the big bang universe the quantum and thermal fluctuations of space-time, radiation, matter distribution must have been highly non-uniform and randomly disordered. Such conditions could have produced localization of these fields. The consequence of such a possibility may need to be included in the proper description of the early universe.

The question we pose, is what happens to the gravitational field for a random distribution of energy and matter? The formulation of this disorder problem in gravitation is rather subtlr. Some of the difficulties arise from the tensor character of the gravity field. Further, the requirements of the principle of equivalence or the symmetry under coordinate transformations also need to be met. Namely, matter at any position influences the geometry at that position; however, the apparent geometry can be transformed away (along a line) by a suitable choice of a free-fall coordinate system. Hence in GR the problem of random distribution of sources has to be posed in a frame independent manner.

Notice, even though over a small region, (~1/g, where g is the determinant of the metric) the effect of any one of the geometric patches can be gauged or transformed away, the global effect of all these contributions is non-zero. This is a key relationship among all non-local phenomena and give rise to Aharonov- Bohm type effects.

5. A Model For Random Gravity

As an example of a disordered gravitational problem, we consider a random perturbation imposed upon the general relativistic background of a stationary, axially symmetric, line mass. This choice of a highly symmetric model is motivated by analytic considerations and computational ease but is not essential for this discussion.

We investigate the effects of disorder by means of a "generating function", A(r,t). This statistical function A(r,t) will introduce disorder through g_{rr} component of the metric tensor and hence into the equations of motion.[9-10] Because, in the absence of disorder the stress-tensor vanishes (except on the axis), in presence of disorder we will require this condition to remain true on the average. That is <stress-tensor> = 0, where < > denotes the ensemble average over all the replicas of the distribution.

We consider null-radial outgoing geodesics. We thus start with a metric given by

$$ds^2 = [1 + \Lambda(r,'t)]\, r'^{\,(2m^2 + 2m)}\, (dr')^2 + r'^{\,(2+2m)}\, d\phi^2 + r'^{\,-2m}\, dz^2 - r'^{\,(2m^2+2m)}\, (dt)^2 \qquad [5.\ 1\ a]$$

where r' is the radial distance from the axis, t is the time coordinate, ϕ is the azimuth angle, z is the axis coordinate, and m is related to the mass parameter (mass per unit length as measured at infinity). We assume that the generating function, A(r,'t), depends only upon r and t and that it is differentiable in these variables. A is partitioned into a non-stochastic and a stochastic part, as follows: [11,12]

$$A(r',t) = A_0 + a(r',t) \qquad [5.\ 1b]$$

The average of A is positive and is represented by A_0. The stochastic part is $a(r',t)$ which is a random function with negative and positive values but of zero mean. The absolute value of $a(r',t)$ is assumed to be smaller than that of A_0. With this stipulation, the matter or source is held positive everywhere although it has a random perturbation above and below its average. Alternatively, we may argue that the mass of the source only makes sense at infinity and as long as the integral of a is less that 1, there is no problem of negative mass.

For any given $A(r',t)$ the terms $1+A_0$ can be rescaled to unity with the new radial distance given by r . In terms of the new rescaled coordinates, r and t, Eq. 1a can be expressed as

$$\begin{aligned} ds^2 = {} & [1 + a(r,t)]\, r^{(2m^2+2m)} (dr)^2 \\ & + r^{(2+2m)} d f^2 + r^{-2m} dz^2 - r^{(2m^2+2m)} (dt)^2 \end{aligned} \qquad [5.\ 1c]$$

In the rest of this paper only these rescaled space time variables r and t will be utilized. For any particular realization, the mixed Einstein tensors are;

$$G_r{}^r = 0,\ G_r{}^t = -\frac{1}{2}\,[a_{,t} \times r^{(-2m^2-2m-1)}],\ G_t{}^t = -\frac{1}{2}\,[a_{,r} \times r^{(-2m^2-2m-1)}],$$

$$G_2{}^2 = \frac{1}{2}\,[a_{,tt} \cdot \frac{1}{2} a_{,t}{}^2 + \frac{m^2}{r} a_{,r}] \times r^{(-2m^2-2m)},\ \text{and}$$

$$G_3{}^3 = [m\, a_{,r} + \frac{1}{2} a_{,r} + \frac{1}{2} a_{,tt}\, r - \frac{1}{4} r\, a_{,t}{}^2 + \frac{1}{2} a_{,r}\, m^2]\, r^{(-2m^2-2m-1)}.$$

For this geometry, these are also the "physical components" of the stress tensor. The null-radial geodesics are given by

$$\frac{d^2r}{d^2t} + \frac{2}{r}(m^2+m)\left(\frac{dr}{dt}\right)^2 + \langle(a_{,r} + a_{,t})\rangle = 0, \qquad [\ 5.\ 2\ ]$$

where τ is an affine parameter.

The ensemble averaging of Eq. 2, is chosen independent of r and t such that

$$\langle G_r{}^t\rangle = \langle G_t{}^t\rangle = 0.$$

We can also make $\langle G_2{}^2\rangle = \langle G_3{}^3\rangle = 0$, provided the generating function satisfies

$$a_{,tt} = \frac{1}{2}(a_{,t})^2 - \frac{m^2}{r}\,a_{,r}. \qquad [5.3]$$

Eq. 5.3 is the choice for the outgoing geodesics. Under these conditions we can choose the ensemble such that $<a_{,r} + a_{,t}> = \alpha$ is a positive constant for $r< r_{critical}$. For $r>r_{critical}$ we take the ensemble average to rapidly approach zero.

6. Localization Parameter for Gravitation

To calculate the effects of disorder a suitable "test" parameter has to be defined. In the electronic problem discussed earlier, Anderson chose the diffusion coefficient as the measure of localization. In the GR-problem, we propose the value of the affine parameter τ along out-going null geodesics to be the fiduciary quantity. The ratio of τ between two events with disorder and the τ for the same two events in absence of disorder is close to unity for no localization. This ratio is larger than one for weak localization and will diverge for strong localization. Similarly we use the differences $\Delta(\tau) = |\tau(\alpha)-\tau(\alpha=0)|$ and $|\Delta(\tau)/\tau(\alpha=0)|$ as measures of localization. $\Delta(\tau)$ is small for extended (non-localized) states and large for the localized states.

The null geodesics, which include both light and gravity waves in this approximation, are given by

$$\frac{d^2r}{d^2\tau} + \frac{2}{r}(m^2+m)\left(\frac{dr}{d\tau}\right)^2 + \alpha = 0. \qquad [6.1]$$

The form of the solution to Eq. 4 depends upon the size of α as compared to r. We can best examine this by first considering the case of $\alpha = 0$. In that case there is no disorder and

$$\tau = \frac{r^p}{p \times c1} + c2 \qquad [6.2]$$

where $p = 1 + 2(m^2 + m)$ and c1 and c2 are arbitrary constants. We will set c1 = 1 and c2 =0 for simplicity.

For $\alpha > 0$, and letting $q = 1 + 4(m^2 + m)$, we define

$$r_0 = \left(\frac{q}{2\alpha}\right)^{1/q}. \qquad [6.3]$$

Then for $r < r_0$,

$$\tau = \frac{1}{q}\left(\frac{q}{2\alpha}\right)^{\frac{q+1}{2q}} \int_{0}^{\left(\frac{r}{r_0}\right)^{q}} \lambda^{\frac{1-q}{2q}} (1-\lambda)^{-\frac{1}{2}} \, d\lambda \qquad [6.4]$$

This integral is just the incomplete Beta function of arguments $[(\frac{q+1}{2q}, \frac{1}{2}]$.
For $x > x_0$,

$$\tau = \frac{1}{q}\left(\frac{q}{2\alpha}\right)^{\frac{q+1}{2q}} \times$$

$$\left[\text{Beta}(\frac{q+1}{2q}, \frac{1}{2}) + \int_{\left(\frac{r_0}{r}\right)^{q}}^{1} \lambda^{-(1+\frac{1}{2q})} (1-\lambda)^{-\frac{1}{2}} \, d\lambda \right]. \qquad [6.5]$$

We have numerically calculated the difference $\Delta(\tau)$ and the behavior of for q = 2 and α from 0.0 to 0.05 in the range of $0 < r < 5$, is shown in Fig. 2. The critical value for $\alpha = 0.04$ is $r_0 = 5$ and for $\alpha = 0.05$ occurs at $r = 4.47$. As the figure shows, the τ needed to reach any given location (r) rapidly becomes larger as the magnitude of disorder (α) is increased.

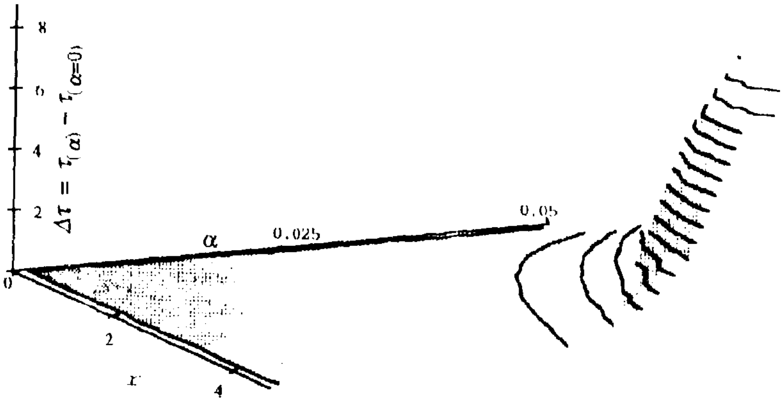


Figure 2. The dependence of (affine parameter)$_{\alpha \neq 0}$ minus (affine parameter)$_{\alpha = 0}$ for $0 \leq \alpha \leq 0.05$ and $0 < r \leq 5.0$.

Conversely, r rapidly decreases for a given τ if $\alpha \neq 0$ as compared to $\alpha = 0$. This would correspond to a retardation of either light or gravitational waves. We have also calculated the value of this difference over an extended range of r for $\alpha = 0.4$. The dependence of $\Delta(\tau)$ for a fixed value of α (α =0.4) and q=2 is shown in Fig. 3

As can been seen from Figs. 2 and 3, a radial null geodesic is retarded by the presence of a random source (as compared to zero disorder). This behavior is reminiscent of the slowing down of localized particles and fields. The electronic and EM -localization properties can be measured in the laboratory; but terrestrial gravitational fields are 40 orders of magnitude weaker. So, GR- localization effects may be observable only in the cosmic scale. However, the non-linear, self coupling not included in the above calculation can amplify these effects. Even for photons at high intensity and strongly scattering non-linearity is known to produce super radiant behavior.

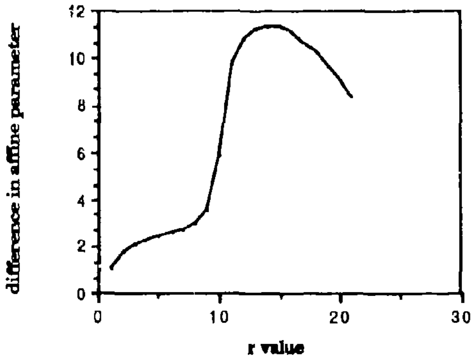


Figure 3. Affine parameter difference for $\alpha = 0.4$, q = 2.

7. Scattering of Gravitational Waves and the Rayleigh Cross-Section

ABefore proceeding further we make a number of observations. All modes (frequencies) are not equally localizable.[5] High frequency short wavelength modes lie in the geometric optics limit and behave ballistically with little interference. For llong wavelength, λ, the gravitational quadrapole[6,7] scattering is dominant and the scattering cross section $\Sigma(\lambda)$ is given by:

$$\Sigma(\lambda) \sim (\lambda^{-6}) \qquad [7.1]$$

This large inverse power law behavior is similar to the well known Rayleigh cross-section in the dipole scattering of light. The gravitational cross-section vanishes more rapidly with an exponent six compared to the fourth power for light. Hence, back-scattering and localization will be strongest over a window of frequency with intermediate values of wavelengths i.e., $\lambda \sim (\Sigma^{1/2})$ $-1 < \xi < r_{\mathrm{criti}}$ $\mathrm{\imath} = r_c$, where ξ is a measure of the spatial correlation length of the distribution function $\mathrm{\imath}(r,t)$.

8. Cosmology of a GR-Localized Universe

In the primordial explosive models for the universe (such as the big-bang), the quantum and thermal fluctuations of space-time and all fields must have been highly non-uniform and randomly disordered at very early epochs. A detailed general relativistic calculation including the non-linear interactions is very complicated, so it will not be attempted here,. Instead we will reason in the big-bang universe disorder may have been sufficient to induce localization of the type discussed above.[4,5]

Under this condition as in any random-walk, time evolution will be determined by the random fluctuations and not by inertia. Propagation will be limited by the diffusion coefficient D(R). For gravitational propagation in a "GR-localized" universe of radius R, D(R) may be obtained by extending the scaling theory results, i.e., [5]

$$D(R) \sim c\Sigma\{r_c)^{-1} + (R)^{-1}\} \qquad [8.1]$$

In Eq.[8. 9], c is the speed of light and $R = S(t)R_0$, where S(t) is the cosmological scaling factor.[10.] The Einstein relation[8] for D, and Eq. [8.9] may be combined to include the effects of temperature (T). Hence, the dynamical equation for S(t) is modified. In the Newtonian limit, with D(S) given by Eq. [8.9] it follows that

$$\frac{d<S(t)>}{dt} = \frac{cG}{3}\frac{D}{kT}<S>^{-2}. \qquad [8.2]$$

Eq.[8.10] shows that the rate of cosmological expansion is controlled by the two factors D and T. Paradoxically, an increase in the temperature slows down the expansion. This is a manifestation of the fluctuation-dissipation theorem and is physically due to the increase in the frequency of scattering at higher temperature.[5] Let us investigate two time regions. At very early times when $R<r_c$, Eq.[8.10] goes to the limit

$$\frac{d<S(t)>}{dt} - \frac{cG}{3}\frac{c\Sigma}{R_0kT}<S>^{-3} = 0.$$

During this period the expansion rate $<S> \sim t^{1/\gamma}$, with $\gamma = 4$. At later times S increases and $R > r_c$ and the rate of expansion is given by

$$\frac{d<S(t)>}{dt} - \frac{cG}{3}\frac{c\Sigma}{r_ckT}<S>^{-2} = 0$$

This epoch has anexpansion rate with $\gamma = 3$. Both of these values (4 and 3) of the exponent γ are higher than the value (2) of the classical Brownian motion expansion rate exponent. Either case represents "critical slowing down" reminiscent of critical behavior observed at phase transitions.

These rates are much slower than that predicted in the conventional non-disordered cosmology. Such gradual expansion must have held the hot primordial matter close together for a period longer than has been previously anticipated. This retardation effectively increases the strength of the gravitational interaction G in the primordial medium. Strong interaction could have helped the matter condense and precipitate at randomly distributed nucleation sites. Such precipitation in condensation cells where the effective G was large may have created the structures and non-uniform distribution of the matter presently observed in the universe.

Acknowledgments

We acknowledge useful conversations with Yakir Aharonov. This work is supported in part by grants from the University of South Carolina and the Office of Naval Research.

References

1. Freeman J. Dyson, "The Dynamics of a Disordered Linear Chain", **Phys. Rev. 92** 1331 (1953).

2. P. W. Anderson, **Phys. Rev. 109**, 1492 (1958); "Local Moments and Localized States", **Rev. of Mod. Phys., 50** , 191 (1978).

3. Nevill F. Mott, "Electrons in Glass"; **Rev. of Mod. Phys. 50**, 203 (1978); Nevill F. Mott and E. A. Davis "*Electronic Processes in Non Crystalline Materials*" 2nd. Ed. Clarendon Press, Oxford (1979).

4 R. Balian, R. Maynard and A. Toulouse, Eds., "*Ill-Condensed Matter*", North-Holland, New York (1980).

5. Sajeev John, "Localization of Light", ***Phys. Today* 44**, 32 (1991). E.N. Economu and M.H. Cohen, **Phys. Rev. Lett. 22**, 1065 (1970). T. Datta, "Electronic Prperties of Non-Crystalline Materials"***Appld. Phys. Comm*, 3**,1-31 (1983).

6. Charles W. Misner, Kip S. Thorne and John Archibald Wheeler , *Gravitation*; Freeman, San Francisco (1971).

7. Steven Weinberg, *Gravitation and Cosmology*; John Wiley & Sons, New York (1972).

8. S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy"; Reprinted from Rev. of Mod. Phys., *15 Noise and Stochastic Processes*, Nelson Wax, ed. Dover, New York (1953) . G.E. Uhlenbeck and L.S. Ornsteien, "On the theory of the Brownian Motion" ibid., 93.

9. T. Datta and John L. Safko, Unpublished results, 1992.

10. D.W. Sciama, *Modern Cosmology* ; Cambriidge Univ,Press,London (1975). Jayant Narlikar, *The Structure of The Universe;* Oxford Univ. Press,New York(1980).

SECTION 6

QUANTUM NON-LOCALITY AND GEOMETRY

THE SUPERPOSITION PRINCIPLE IN MULTI-PARTICLE SYSTEMS

DANIEL M. GREENBERGER
Department of Physics, City College of the City University of New York,
New York, New York 10031

O. W. GREENBERG
Center for Theoretical Physics, Department of Physics, University of Maryland
College Park, MD 20742-4111, USA

and

T. V. GREENBERGEST
Department of Physics, Southern Methodist University, Dallas, TX 75275, USA

ABSTRACT

We point out that in two-particle Down-Conversion experiments the photon pair is created in an entangled state, which leads to multiparticle interference, as seen in coincidence counting between the pairs. We note that this type of coherence is different from, and incompatible with, standard single-particle interference.

First, I would lik[illegible]y homage to Yakir Aharonov, and congratulate him both on his birthday and on thi[illegible]nderful symposium, which is worthy of the occasion. I had a friend who, on his sixtieth birthday toasted himself with, "Well, I'm halfway there!", and I wish to Yakir the same optimism and promise of continuing youth and productivity it implies.

This talk should be seen as a sort of addendum to the talks by Profs. Ray Chiao and Anton Zeilinger at this symposium. They both talked of the wonderful experiments that have been performed, and are yet to come, with parametric down-conversion.[1] I am merely going to examine the production process in the light of the superposition principle, in order to show that superposition with many particles is even richer than it is for one particle.

In the down-conversion process, a single photon hits a non-linear crystal, and two photons emerge. Inside the crystal, energy and momentum are conserved, which correlates the momentum of the product photons. By placing a screen on the side of the crystal where the photons are emerging, one can put pinholes so placed as to guarantee that the emerging photons have the correct momenta to satisfy the conservation criterion. We shall make a simple model of this process that ignores the dynamic details, and only considers the wavelike features consistent with superposition and the uncertaintly principle.[2,3]

To this end, consider the total momentum of the photons as zero, to within the uncertainty principle, as guaranteed by the size of the crystal and the placement of the pinholes. So the two photons will emerge on opposite sides of the crystal, each pass through two pinholes, and then impinge on a screen, as shown in Fig. (1). In an actual experiment, the screen is usually replaced by a beam-splitter, where photons are shunted into one of two particle detectors. But this is a complication that does not affect the physics we are considering.

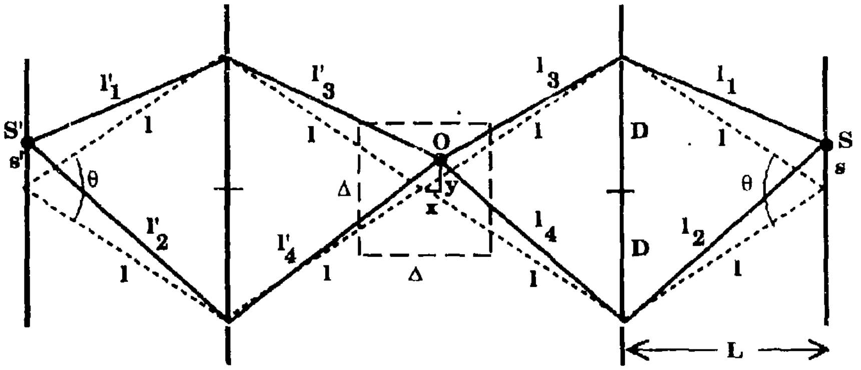


Fig. (1). A simple model for two-particle down-conversion.
The perfect geometry to the center of the pattern is indicated by dashed lines. The solid lines represent the actual paths from O to S and S'. The position of O is given by (x,y), and of S and S' by s and s'.

The amplitude a(O,S) for photon 1 to go from the source point O to the screen point S will be proportional to

$$a(O,S) \sim (e^{ik(l_1+l_3)} + e^{ik(l_1+l_4)}).$$

From the diagram, we see that

$$l_1 = (D-s)^2 + L^2 \sim l - \theta s/2,\ l_2 \sim l + \theta s/2,$$

$$l_3 \sim l - \theta y/2 - x,\ l_4 \sim l + \theta y/2 - x.$$

Thus,

$$a(O,S) \sim e^{ik(2l-x)} \cos\tfrac{\pi\theta}{\lambda}(s+y),$$

since $k = 2\pi/\lambda$. Similarly, the amplitude a(O,S') for photon 2 to go from O to the point S' on the other screen will be

$$a(O,S') \sim (e^{ik(l_1'+l_3')} + e^{ik(l_1'+l_4')})$$

$$\sim e^{ik(2l+x)} \cos\tfrac{\pi\theta}{\lambda}(s'+y),$$

since

$$l_1' \sim l - \theta s'/2,\ l_2' \sim l + \theta s'/2,$$

$$l_3' \sim l - \theta y/2 + x,\ l_4' \sim l + \theta y/2 + x.$$

This leads to a result for the amplitude to detect both a photon at S and one at S' in coincidence. It is

$$a(S,S) = \frac{1}{\Delta^2}\int_{-\Delta/2}^{\Delta/2}\int_{-\Delta/2}^{\Delta/2} dx\,dy \cos\tfrac{\pi\theta}{\lambda}(s+y)\cos\tfrac{\pi\theta}{\lambda}(s'+y)$$
$$\sim \cos\tfrac{\pi\theta}{\lambda}s\,\cos\tfrac{\pi\theta}{\lambda}s', \qquad \Delta << \lambda/\theta,$$
$$\sim \cos\tfrac{\pi\theta}{\lambda}(s-s'), \qquad \Delta >> \lambda/\theta.$$

The limit $\Delta << \lambda/\theta$ represents a point-like source (much smaller than the fringe spacing at the screen), and is the usual criterion that must be satisfied to see fringes from a single source. From a source larger than this, the fringes at the screen will wash out. In our case, such a small source will create two independent diffraction patterns at the two screens, so that we will have each photon leaving the source and acting independently of the other. However in the limit $\Delta >> \lambda/\theta$, for a diffuse source, one sees a correlated two-particle interference pattern. One sees the two-particle pattern by having one detector at point S at one screen and the other detector at point S' at the other screen. If one sits at point S and varies S', one will have a sinusoidal counting rate, $P(s,s') \sim |a(S,S')|^2$; varying with s', which has a 100% visibility. In this case if one counts a single particle, say at S, without simultaneously detecting the other particle, one will find no diffraction pattern. The probability of detecting a single particle at S is independent of s, since in this case

$$P(s) = \int ds' |a(S,S')|^2 = const.$$

The interference only occurs in two-particle coincidences.

There is a simple physical reason for this result, which is truly quantum-mechanical. Normally, for single particle diffraction from a source, through a pair of pinholes to a screen, one knows the position of the source accurately. This is due either to the source itself being small, or through the use of lenses to position it at infinity, so that the wave fronts at the two slits are correlated. This emission from an effective point-source produces a diffraction pattern at the viewing screen. In our case of two-particle diffraction, neither particle by itself forms a diffraction pattern, but the very fact that one of the particles strikes the screen at a particular point S implies a certain knowledge about the source. It says that the source is most probably located at a position O such that the difference in the two path lengths from O to S is a multiple of the wavelength, so that they are in phase. Thus the very fact that one particle lands at S sets up a virtual lattice of probable positions for the source. This in turn will produce a diffraction pattern for the other particle at the other screen. So although either particle can land anywhere, once one has done so, it is closely correlated to where the other particle is likely to land.

One can also get some insight into the process by considering the emission in momentum space. For a large source one has Δ for the transverse size of the source (in the y direction). This implies that $\delta p_y \sim h/\Delta$, and since for a large source, $\Delta >> \lambda/\theta$, this says that the angular spread of the emitted photon is $\varphi \sim \delta p_y / p \sim h/\Delta + h/\lambda << \theta$. So the angular spread is too narrow to encompass both slits. The photon goes through one slit or the other and there can be no interference between the two paths. On the other hand this very narrowness guarantees that if photon 1 goes through the upper slit, then photon 2 will go through the lower slit, and vice versa. So the two photons are correlated and the wave function will be

$$\psi \sim (|k\rangle_1 |-k\rangle_2 + |-k\rangle_1 |k\rangle_2),$$

an entangled state. On the other hand if the source is very small, we will have $\varphi >> \theta$, so that one cannot tell which slit photon 1 will enter and there will be interference between the two slits. This same lack of definition of the beam guarantees that one also cannot correlate the slit for photon 1 with that for photon 2, and so one will get two independent one-particle interference patterns.

The lesson we learn from all this is that there is a complementarity between one-and two-particle interference.[4] If one is present, it excludes the possibility of the other. One either has correlated coincidence counts between the particles and an entangled state to describe them, or one has independent interference patterns for the separate particles, and one describes them by a product state.

As a final note, we mention a three-particle source. The importance of three-particle sources are that they are needed to create GHZ states, to test Bell's Theorem without inequalities.[5] Here three particles are emitted in a plane. If they are identical particles, momentum and energy conservation says that they will be emitted at 120° with respect to each other, each with the same energy. Again we will assume they are emitted at some point O within a source of size Δ in each dimension, and they land the points S, S', S'', at their respective screens. The source is again taken to have coordinates x, y, as in Fig. (2), and the amplitude for landing at S, S' S'', will be

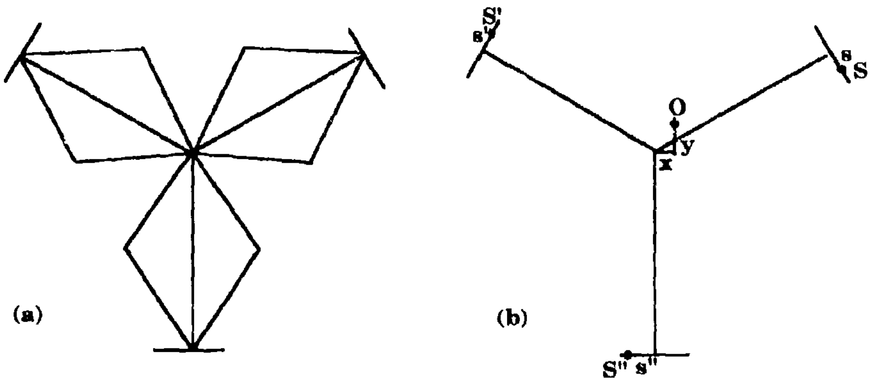


Fig. (2). The three-particle interferometer.
(a) The perfect geometry; (b) The actual source point, O, and screen detection points S, S', and S'', defined similarly as in Fig. (1).

$$a(S,S',S'') \sim \frac{1}{\Delta^2}\int_{-\Delta/2}^{\Delta/2}\int_{-\Delta/2}^{\Delta/2} dx\,dy\, a(O,S)a(O,S')a(O,S'')$$

$$\sim \cos\tfrac{\pi\theta}{\lambda}(s+s'+s''), \qquad \Delta >> \lambda/\theta,$$

$$\sim \cos\tfrac{\pi\theta}{\lambda}s\,\cos\tfrac{\pi\theta}{\lambda}s'\,\cos\tfrac{\pi\theta}{\lambda}s'', \quad \Delta << \lambda/\theta.$$

So once again we see that for large Δ, we have an entangled state, this time between the three particles, and in this case there is neither single particle, nor even two-particle, coherence. Also for small Δ, as before, the three particles produce independent product amplitudes showing one-particle interference, but no multi-particle correlations.

This work was supported in part from grant #92-13964 from the National Science Foundation.

REFERENCES

1. Down-conversion as a process was discovered by D. C. Burnham and D. L. Weinberg, Phys. Rev. Lett. **25**, 84-87 (1970), but it was not used in two-particle interferometry until the middle 1980's. The first experiments were C. O. Alley and Y. H. Shih, Proc. of the *Second International Symposium of Foundations of Quantum Mechanics in the Light of New Technology*, M. Namiki, *et al.*, eds., Physical Society of Japan, Tokyo, 1986, p. 47; R. Ghosh and L. Mandel, Phys. Rev. Lett. **59**, 1903-5 (1987).
2. A more general overview of the process we shall describe will be given in an article in Physics Today, (to be published, July or August 1993) by D. M. Greenberger, M. A. Horne, and A. Zeilinger.
3. The specific model we discuss for the two-particle casewill be treated in greater detail in G. Jaeger, M. A. Horne, and A. Shimony (to be published).
4. Many experiments have been done to confirm this conclusion. However the only one that uses a configuration equivalent to the four slit scheme for the two photons was a Bell-type experiment by J. G. Rarity and P. R. Tapster, Phys. Rev. Lett. **64**, 2495-2498 (1990).
5. For the Bell Theorem without inequalities, see, for example, D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, Amer. J. Physics **58**, 1131 (1990). This paper also explains the relevance of the three-particle interferometer we have considered. For an example using spin, see N. D. Mermin, Physics Today **43**(6), 9 (1990): Amer. J. Physics **58**, 731 (1990).

Non-Locality and Objectivity in Quantum State Reduction

Roger Penrose
Mathematical Institute, Oxford, UK

Abstract

An example of quantum non-locality is presented ("magic dodecahedra") which illustrates Bell's theorem without probabilities. A scheme is then put forward for the objective reduction of the quantum state vector when too large displacements of mass are involved in a superposition between two quantum states. In this scheme, the reduction time is roughly the reciprocal of the gravitational self-energy of the difference of the two mass distributions, measured in absolute units.

It is a pleasure to be able to pay my respects to Yakir Aharonov in honour of his 60th birthday. I shall briefly describe two ideas that have to do with that subject - *quantum mechanics* - which has engaged so much of his attention, and to which he has made so many surprising and profound contributions. The first is an example that illustrates one of the theory's most puzzling features: quantum (Bell) non-locality, without probabilities. The second represents a new angle on the measurement problem.

1 Magic Dodecahedra

I have described this non locality example several times elsewhere [1,2,3] so it will not be necessary to give more than a very brief outline of what is involved. The system consists of two atoms of spin 3/2 which are initially produced in a combined state of spin 0 and then slowly separated to a great distance from one another without disturbing their individual spins. Measurements are subsequently made on the two atoms individually, each measurement being of a particular yes/no kind and corresponds to one of 20 possible directions in space: those which are represented by the vertices of a regular dodecahedron, as measured out from the centre. Thus, we imagine two widely separated but parallel-oriented regular dodecahedra (which, for dramatic effect, we can be imagined as being here on earth and on a planet orbiting α-Centuri, respectively), each of which has a spin 3/2 atom at its centre. Each measurement would be defined by choosing one of the vertices of one dodecahedron and ascertaining whether the amount of its central atom's spin in that direction - i.e. the m-value, in that direction - has the particular value 1/2. If this is found to be so, then this is the answer "yes"; and one envisages that a bell rings, indicating that the measurements on that particular atom have come to an end. If the value 1/2 is not obtained ("no"), then the three possible $m\neq 1/2$ states (namely 3/2, -1/2, -3/2) are combined without disturbing the phase relations between them, and the measurement is repeated in some other direction.

For example, we could envisage performing the measurement with a Stern-Gerlach type of apparatus, oriented appropriately in the chosen direction, and where only *one* of the four different beams (the one corresponding to $m=1/2$) is examined, yielding the "yes" answer (bell ringing) if the atom is found in that beam. Otherwise, by appropriate reversing of the magnetic fields, the three remaining beams are brought together without disturbing their relative phases. An exactly similar spin measurement is then performed in some other direction, corresponding to a different vertex of the dodecahadron, and so on.

There are just two different properties that we shall need, concerning the results of the joint measurements of this type that can be performed on the two atoms - by myself here on earth and by my colleague Alfie St.Uri, on α-Centuri. These concern sequences of measurements of the following type. One of the vertices of the dodecahedron is singled out - call this vertex the SELECTED one - and measurements are performed corresponding to the three vertices of the dodecahedron that are *adjacent* to the SELECTED one (but not in the direction of the SELECTED one itself). It may be ascertained that the "yes" eigenstates of these three measurements are all *orthogonal* to one another, and it follows that the three measurements necessarily *commute* - so it makes no difference in which order the three are performed. We deduce the first of the two properties that we shall need:

> (1) If Uri and I happen to SELECT *diametrically opposite* vertices on our respective dodecahedra, then the bell rings for one of my measurements if and only if it rings for Uri's diametrically opposite measurement, this being irrespective of whether it rings on the first, second, or third of the measurements adjacent to our SELECTED vertices.

The second property is a little harder to ascertain, although this can be done without further explicit calculation (cf. ref. 1 for details):

> (2) If Uri and I happen to SELECT *corresponding* vertices on our respective dodecahedra, then the bell must ring for at least one of the six measurements that we propose to make.

If we are to assume that the bell-ringings are determined according to some kind of local hidden variable theory - or, simply, that what happens on α-Centuri is completely independent, in the ordinary classical way, of the measurements that I choose to make here on earth, then we can quickly deduce three things concerning the results of measurements on my own dodecahedron alone:

> (a) Each vertex of my dodecahedron is preassigned as either a *bell ringer* (colour it WHITE) or as *silent* (colour it BLACK), irrespective of the ordering in which the measurements are made adjacent to any SELECTED vertex.
> (b) No two next-to-adjacent vertices can be both bell-ringers (WHITE).
> (c) The six vertices adjacent to one or other of a pair of opposite vertices cannot be all silent (BLACK).

It is a nice combinatorial exercise to show that no colouring of the vertices of a dodecahedron WHITE or BLACK is possible, according to the rules (b) and (c). This shows that the assumption of classical independence between the atom here on earth and the atom on α-Centuri must be false - assuming that the expectations of standard quantum theory are maintained. In Einstein's terminology, there is a "spooky action at a distance" between the results of the measurements on the spin 3/2 atoms that Uri and I might choose to make. For earlier examples of Bell non-locality without probabilities see[4–9], and the review article by Brown[10]; also [11–13] for results that can be adapted to give non-local examples of this nature.

2 The Role of Gravity in Quantum State Reduction

I have frequently argued the case that the phenomenon of state-vector reduction must be a *real* physical effect of some kind, and not just an illusion, or a property of conscious observers, or just some tricky matter of finding the right "interpretation" of the quantum formalism. Moreover, I have maintained that the physics that is involved must be something in which the effects of *gravity* are crucial (cf. also [14–17]). Of course, it is clear that there are many differing viewpoints with regard to this phenomenon, and I shall certainly make no serious attempt to convert anybody. The motivations that underlie my own approach are various, but I believe that a number of independent arguments can be given in favour of a fundamental role for gravity in quantum state reduction .

For me, one of the strongest comes from the study of the *space-time singularities* in the big bang and black holes. As a fundamental ingredient to the second law of thermodynamics, it is necessary that the big bang's singularity must have been enormously constrained - to such an extraordinary precision that only one part in (at least) $10^{10^{123}}$ of the available phase space was made use of. (Very likely, the precision is considerably greater than this, depending upon the actual baryon content of the universe the precision being infinite for a spatially infinite universe.

This figure is calculated on the basis of the Bekenstein- Hawking formula for black hole entropy, assuming a total baryon content of about 10^{80}. The necessity of this kind of precision is not removed by inflation[18].) It is this enormous constraint on the gravitational degrees of freedom in the early universe, together with the fact that the singularities of black holes - or of an all-embracing big crunch - seem to be subject to no constraint at all, that gives us the powerful second law of thermodynamics in the form that we know it. The structure of space-time singularities is generally accepted to be a quantum gravity effect - or at least an effect of whatever the correct union of quantum theory with gravitational theory might be. Since the initial and final types of singularity seem to need to have such grossly different structures, this strongly indicates that whatever this quantum-gravity union might turn out to be, it must be a time-asymmetric theory. The indications are, therefore, that something more than just the standard time-symmetric procedures of unitary evolution must be involved; the time *a*symmetric phenomenon of quantum state reduction must also be part of this entire unified picture.

It is possible to be more explicit about this link between the time-asymmetry of space-time singularities and that of the reduction procedure[19,20], but I have no wish to repeat the arguments in detail here. The essential point is that the complicated high-entropy singularities of gravitational collapse serve to "absorb information", causing an effective *reduction* in phase-space volume. Over the totality of all possible states, this must be precisely balanced by a corresponding effective *increase* in phase-space volume that results from an indeterminacy in the evolution of physical systems. This indeterminacy is argued to be that which is arises in state-vector reduction. (The phase-space volume increases because, in effect, when the state gets reduced there are several different alternative outputs for each input; whereas, given the output, there is generally only one plausible input that need be considered.) It is this necessary balance between these two seemingly disparate parts of physics that tells us that these two parts of physics must actually be one and the same. Thus, not only is the structure of space time singularities - and consequently also the second law of thermodynamics - a quantum gravity effect, but so also must be the very process of quantum state reduction.

I have argued earlier that state reduction should be something that comes about when space-times would have to be superposed which differ "too much" from one another, in the sense that the difference between the space-times is of the order of "one graviton" or more (so nature abhors superpositions between sufficiently different space-time geometries). I have now come to the conclusion that we should not regard this measure of difference as representing something absolute - for which quantum linear superpositions would be forbidden whenever this value is exceeded. Rather, we should consider that there is a *rate* at which reduction occurs, this rate being large for space-times that differ by a large amount, and small, when the space-times do not differ much. Thus there is to be an instability involved in space-time superpositions, giving a kind of *half-life* for the superposed state, that is of the order of the reciprocal of the appropriate measure of the difference between the superposed states.

What is this measure of difference? In a recent article[21], I gave some rather tenuous motivations for a measure of difference between the two weak quasi-static gravitational fields that are associated with two different Newtonian mass distributions. This can be reformulated as the *gravitational self-energy* of the *difference* between the two mass distributions. The present proposal, then, is to take this self-energy E, measured in Planckian units (i.e. the *absolute* units for which $G = c = \hbar = 1$). Then E^{-1} gives something, in Planckian units, that is of the order of the *time* that the superposition persists before it reduces to that given by either one mass distribution or the other. That is to say, E^{-1} is roughly a half-life for the superposed state to decay into one state or the other.

There is a particular advantage in a viewpoint of this nature that is not shared by most other proposals for "realistic" quantum state reduction (such as that of Ghirardi, Rimini, and Weber[22]). In the case of an unstable particle, there is always an uncertainty in the mass of the particle, this uncertainty (in units for which $\hbar = c = 1$) being of the order of the reciprocal of the lifetime. Thus, for any state-reduction process of the general kind that I am proposing here, we expect some kind of mass-energy uncertainty that is inherent in the superposed state. With the present proposal, this uncertainty would have to be of the same order as the self-energy in the gravitational field of the difference between the two mass distribution under consideration. This self-energy, according to classical general relativity, is not well defined - or, at least, it is not localizable - in a coordinate independent way.

In classical general relativity, this is an inherent feature of the theory. The quantity T_{ab} that occurs in Einstein's equation $R_{ab} - 1/2Rg_{ab} = 8\pi T_{ab}$ describes all the energy of *matter*, but it does not directly take into account the energy in the gravitational field. That energy is non-local, and cannot be meaningfully assigned a local measure of density. Thus, there is no tensor quantity, independent of coordinate choice, which describes this energy. Nevertheless, gravitational field energy is "real", in the sense that it must be taken into account in physical processes, such as the (positive) energy that is carried away in the form of gravitational radiation from a double neutron star system, or the (negative) contribution to the total mass of a celestial body, such as Jupiter, owing to its gravitational self-energy. However, the mass-energy in gravitation is a fundamentally slippery quantity, which cannot be meaningfully localized.

A feature of the present proposal for quantum state reduction is to take advantage of this slipperiness in order to evade an energy problem that seems to be an essential feature of any model of state reduction in which that process is taken to be a "real" phenomenon. Basically, if "quantum jumps" are taken to be *real*, then the mass-energy distribution in a system undergoes local violations of energy conservation when the jumps occur. In the original reduction scheme due to Ghirardi, Rimini, and Weber[22] (GRW), for example, there is a small energy violation involved in the "hits" that effect the reduction process. One of the physical ingredients of the present scheme is that there is the potential possibility of dovetailing one of these energy problems with the other - that of classical general relativity with that of quantum state reduction - so that a consistent overall scheme may be obtained. At the time of writing, however, I have not worked out the details of how this dovetailing is fully to take place. It should be pointed out, moreover, that the gravitational variant of the GRW scheme put forward by Diósi[17] has considerably more serious energy problems than the original GRW scheme - to the extent that it is in gross conflict with observation, as was pointed out by Ghirardi, Grassi, and Rimini[23]. These three authors suggested a modification that removed this observational conflict, but at the expense of introducing an *ad hoc* parameter that was not present in Diósi's proposal. The present proposal differs from that of Diósi, though it has a number of features in common with it. More details will need to be sorted out before it can be ascertained whether, within the scheme of ideas that I am setting forth here, one can construct a detailed proposal that is completely free of such energy problems.

In the meantime, we can at least examine the orders of magnitude that arise with the present scheme. Let us first take note the values of some of the standard physical units in terms of the dimensionless Planckian ones (for which $G = c = \hbar = 1$):

$$\begin{gathered}\text{second} = 1.9 \times 10^{43},\ \text{day} = 1.6 \times 10^{46},\ \text{year} = 5.9 \times 10^{50},\\ \text{metre} = 6.3 \times 10^{34},\ \text{cm} = 6.3 \times 10^{32},\ \text{micron} = 6.3 \times 10^{28},\\ \text{radius of nucleon} = 10^{19},\ \text{mass of nucleon} = 10^{-19},\\ \text{gram} = 4.6 \times 10^{4},\ \text{erg} = 5.2 \times 10^{-17},\ \text{degree Kelvin} = 4 \times 10^{-33},\\ \text{density of water} = 2 \times 10^{-94}\ .\end{gathered}$$

If we consider a uniform sphere of radius a and mass m whose state is gradually evolved into a superposed state of two different locations, separated from one another by a distance comparable with their radius a, then we find a gravitational self energy E, for the difference between the two mass distributions, which is the order of

$$E = m^2/a.$$

Thus, according to the proposal I am putting forward, this superposed state is unstable, and would decay into the state in which the sphere is *either* in one location *or* in the other, in a time (half life) that is of the general order of

$$T = a/m^2.$$

In terms of the density ρ (assumed uniform) and radius a of the sphere, this is very roughly

$$T = 1/(10\rho^2 a^5).$$

Now consider various examples. For a nucleon, assuming its radius to be its Compton radius

$$\begin{aligned} T &= 10^{19}/10^{-38} = 10^{57} \\ &> \text{million years.} \end{aligned}$$

Since ordinary laboratory experiments take place in time scales that are much less than this, there is – according to the present scheme – no danger of any discrepancy with the predictions of standard quantum theory for a quantum system consisting of just a few nuclear particles. In particular, the results of neutron interference experiments are not contradicted. For a droplet of water, of radius a, we find, approximately,

$$T = 10^{189} a^{-5}$$

so we get, very roughly,

$$\begin{aligned} T &= 10 \text{ days}, && \text{if} \quad a = 10^{-5} \text{ cm} \\ T &= 10^{-1} \text{ seconds}, && \text{if} \quad a = 10^{-4} \text{ cm} = 1 \text{ micron} \\ T &= 10^{-6} \text{ seconds}, && \text{if} \quad a = 10^{-3} \text{ cm.} \end{aligned}$$

Thus we see that, in a sense, a "turnover" from quantum to classical behaviour occurs at roughly a micron's scale. These figures, and other related ones, are not at all unreasonable. They do not seem to contradict anything obvious about quantum or classical behaviour.

In making this statement, I am taking into account two factors that I have not mentioned so far. The first is that in the above estimates I have treated the bodies as *uniform* objects, and not as composed of individual atomic particles. In fact, in any ordinary situation of a superposition between two mass distributions, one of which is a rigid translation of the other, the granular (atomic) nature of the distributions turns out not to be important for the calculation of the reduction rate. But we can also consider a *different* type of situation: where we have a superposition of two mass distributions which do not differ macroscopically at all. Instead, it is now to be their submicroscopic constituents that are located differently in the two states. Depending upon the nature and the amount of movement of these constituents, we find in this situation that, for an equal reduction rate, the total size of the material that is involved in the superposition would tend to be somewhat larger – but not enormously larger – than that considered above.

Secondly, these considerations are important when reduction occurs because the *environment* becomes entangled with the quantum system that is under study. Indeed, according to the present proposal, in any practical situation that one can easily envisage at the moment, it would indeed be the disturbed environment that effects the reduction. Thus we obtain nothing different from the conventional picture of state-vector reduction, in which it is the "decoherence" caused by the environment that causes the reduction - except that now the reduction must be considered as a real physical effect, not just something that takes place "for all practical purposes" (John Bell's "FAPP"). It would, no doubt, take a delicately organized experimental set-up to detect any differences between the present proposal and conventional quantum mechanics. Nevertheless, differences would be detectable in principle. One would need to arrange things so that some "large" quantum system can remain isolated from its surroundings for sufficiently long that, according to the present scheme, state-reduction should take place spontaneously within that time-scale, entailing a loss of phase coherence. On the other hand, standard quantum mechanics would demand that such coherence would be maintained for as long as the system remains isolated.

Finally, it should be remarked that these considerations leave us a long way from an actual *theory* of gravitationally-induced state-vector reduction. The difficulties of providing a coherent picture of the reduction in accordance with the principles of relativity are well known, and were stressed many years ago by Yakir and his colleagues[24].

Acknowledgements

I am grateful to Abhay Ashtekar, Ted Newman, and Lenny Susskind for illuminating suggestions; I also thank NSF contract PHY-9396246 at Syracuse and Penn State Universities for partial funding assistence.

References

1. R. Penrose *On Bell non-locality without probabilities: some curious geometry*, in *Quantum Relections* (in honour of J.S. Bell), eds. J Ellis and D. Amati (Cambridge Univ. Press, Cambridge 1994 to appear)

2. R. Penrose in *The renaissance of General Relativity* (in honour of D.W. Sciama), eds. G. Ellis, A. Lanza and I. Miller (Cambridge Univ. Press, Cambridge 1993)

3. J. Zimba and R. Penrose *Stud. Hist. Phil. Sci.* (1993)

4. P. Haywood and M.L.G. Redhead *Found. Phys.* 13 (1983) 481-99

5. A. Stairs, in *Philos. of Sci.* 50 (1983) 578-602

6. H.R. Brown and G. Svetlichny *Found. Phys.* 20, (1990) 1379- 87.

7. D.M. Greenberger, M.A. Horne and A. Zeilinger *Bell's Theorem, Quantum Theory, and Conceptions of the Universe*, ed M. Kafatos (Kluwer Academic, Dordrecht, The Netherlands 1989), pp. 73-76.

8. D.M. Greenberger, M.A. Horne, A. Shimony and A. Zeilinger *Am. J. Phys.* 58, (1990) 1131-43.

9. R.K. Clifton, M.L.G. Redhead and J.N. Butterfield *Found. Phys.* 21, (1991) 149-84; errata, *Found. Phys. Lett.* 4, (1991).

10. H.R. Brown in *Bell's Theorem and Theorem and the Foundations of Physics* eds. A. Van der Merwe and F. Selleri (World Scientific, Singapore 1993)

11. J.S. Bell *Revs. Mod. Phys.* 38, (1966) 447-52.

12. S. Kochen and E.P. Specker *J. Math. Mech.* 17 (1967) 59-88.

13. A. Peres *J. Phys.* A24 (1991) L175-8.

14. F. Károlyházy *Nuovo Cim.* A42, (1966), 390

15. A.B. Komar *Int. J. Theor. Phys.* 2, (1969) 157-60.

16. F. Károlyházy, A. Frenkel and B. Lukács in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford 1986).

17. L. Diósi *Phys. Rev.* A40, 1165 (1989).

18. R. Penrose in *Fourteenth Texas Symposium on Relativistic Astrophysics* ed. E.J. Fenyves *N.Y. Acad. Sci., New York* 57 (1989) 249- 64.

19. R. Penrose *Quantum Concepts in Space and Time* eds. R.Penrose and C.J. Isham, Oxford University Press 1986), pp. 129 146

20. R. Penrose *The Emperor's New Mind* (Oxford University Press (1989))

21. R. Penrose, in *General Relativity and Gravitation 1992: Part I, Plenary Lectures* eds. R.J. Gleiser, C. Kozameh and O.M. Moreschi (I.O.P. Publ. 1993) pp. 179-89.

22. G.C. Ghirardi, A. Rimini and T. Weber *Phys. Rev.*, D34, (1986) 470.

23. G.C. Ghirardi, R. Grassi and A. Rimini *Phys. Rev.* A42 (1990) 1057.

24. Y. Aharonov and D.Z. Albert *Phys. Rev.* D24 (1981) 359.

A NON-POLARIZATION EPR EXPERIMENT : OBSERVATION OF HIGH-VISIBILITY FRANSON INTERFERENCE FRINGES

Raymond Y. Chiao, Paul G. Kwiat, and Aephraim M. Steinberg
Department of Physics, University of California, Berkeley, CA 94720, U. S. A.

One of us (R. Y. C.) will review a series of experiments recently conducted at Berkeley, including the "quantum eraser" and a "dispersion cancellation" experiment, and culminating in the Franson experiment, in which a violation of a Bell's inequality for energy and time by more than 16 standard deviations is implied. We conclude that, unlike Yakir, photons do not possess well defined birthdays.

1. Introduction

In this talk, I shall briefly review the Einstein-Podolsky-Rosen (EPR) "paradox" [1], and then describe some of our recent experiments at Berkeley in light of this so-called paradox: (1) the "quantum eraser" [2], (2) a "dispersion-cancellation" effect in two-photon interference[3], and (3) the Franson experiment[4, 5], which involves the nonlocal interference of photon pairs. Let me begin by stating my belief that there is no true paradox to EPR, since there are no genuine contradictions, either internally in logic, nor externally with experiment. (Perhaps a better name would have been "the EPR effect.")

The EPR effect involves the interference of two spatially separated particles which are generated by a decay from a common source S in the following geometry:

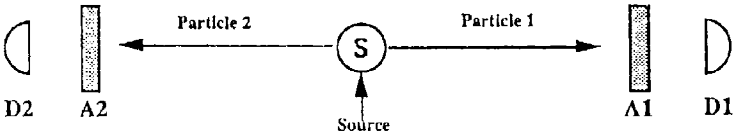


Figure 1. EPR experiment

The two particles are measured by means of analyzers (A1, A2) and detectors (D1, D2). In the Bohm version of the EPR effect[6], for example, a spin-0 particle decays into two spin-1/2 particles in a singlet state

$$|\text{Singlet}\rangle = \frac{1}{\sqrt{2}}\{|\uparrow_1\rangle|\downarrow_2\rangle - |\uparrow_2\rangle|\downarrow_1\rangle\}.$$

The analyzers A1 and A2 are Stern-Ge polarizers. Optical versions of this experiment performed by Freedman and Clauser, an by Aspect *et al*, used photons in place of the spin-1/2 particles, and linear polarizers (P1, P2) in place of Stern-Gerlach polarizers[7, 8].

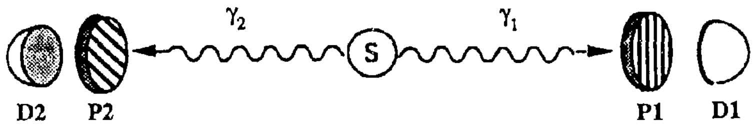


Figure 2. Optical version of EPR experiment

The coincidence count rate as a function of the relative angle between polarizers P1 and P2 is a measure of the correlated behavior of the two separated particles. Bell[9] derived an inequality starting from two very general and seemingly reasonable notions which were introduced by EPR, namely, locality and reality. This inequality is violated by the 100%-visibility sinusoidal fringes predicted by quantum mechanics, however. Most importantly, *experiments* reveal a sinusoidal variation of the coincidence rate in agreement with quantum mechanics, and in violation of this inequality (modulo some reasonable auxiliary assumptions). Therefore, these experiments rule out all local, realistic theories.

Early experiments relied on the correlations of the *polarization* (i.e. an internal degree of freedom) of the particles, whereas the Franson experiment relies on the correlations in the *energy* and the *time of emission* (i.e., external degrees of freedom) of the particles. These *external* degrees of freedom are very similar to the momentum and position of the particles considered in the original EPR paper. Since the predictions of quantum mechanics are so strange, it is critical to investigate them for these external degrees of freedom as well as for the internal ones. Rarity and Tapster have already done so for momentum and position[10]. We have recently done so with energy and time.

2. Entangled states

Schrödinger[11], in response to the EPR paper, pointed out that at the heart of these nonlocal effects is what he called "entangled states" in quantum mechanics, i.e. coherent sums of product states which are *nonfactorizable*. For if a two-particle wavefunction were factorizable,

$$\psi(x_1, x_2) = \chi(x_1)\chi(x_2)$$

then the probability of joint detection would also factorize,

$$|\psi(x_1, x_2)|^2 = |\chi(x_1)|^2|\chi(x_2)|^2$$

so that the outcomes of two spatially separated measurements are *independent* of one another. In cases where quantum mechanics predicts correlations in the behavior of distantly separated particles, this means that the two-particle state *cannot* be factorized as above. The Bohm singlet state mentioned above is a good example of an entangled state, since it is nonfactorizable, and leads to correlations in polarization measurements on remote particles. Though each particle considered individually is unpolarized, the two particles will *always* have opposite spin projections when measured along the same quantization axis. For different choices of the axes these projections are incompatible observables and therefore cannot have definite values simultaneously. But these correlations persist even if the particles and their analyzers are separated by space-like intervals, implying the existence of nonlocal influences. Another good example of an entangled state is the Slater determinant:

$$\begin{vmatrix} \psi_1(x_1) & \psi_1(x_2) \\ \psi_2(x_1) & \psi_2(x_2) \end{vmatrix} = \psi_1(x_1)\psi_2(x_2) - \psi_1(x_2)\psi_2(x_1)$$

which predicts correlated behavior between separated fermions.

In our experiments, the entangled state we start with is the ***energy-entangled*** state of two photons produced in a two-photon decay process known as parametric fluorescence. The Feynman diagram for this process is

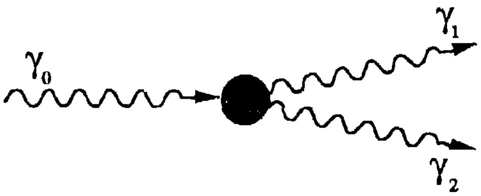


Figure 3. Two-photon decay from one photon

and the state of the two-photon system after decay from a parent photon of a sharp energy E_0 is given by

$$|2 \text{ photons}\rangle = \int_0^\infty\!\!\int_0^\infty dE_1 dE_2\, \delta(E_0 - E_1 - E_2)\, A(E_1, E_2)\, |E_1\rangle\, |E_2\rangle .$$

Instead of a sum, as in the Bohm singlet state, we now have an integral, since energy is a continuous variable. The meaning of this energy-entangled state is that after the measurement of the energy of one photon results in a sharp value E_1, there is an instantaneous collapse to the state

$$|E_1\rangle\, |E_0 - E_1\rangle .$$

This effect has been seen in an earlier experiment[12], in which coincidences are recorded between photon γ_1, which passes through a Fabry-Perot filter (to measure its frequency, and hence its energy, with finite but high resolution), and photon γ_2, which passes through a Michelson interferometer (to measure its width); when photon γ_1 is detected after the narrow-band filter, photon γ_2 collapses into a wave packet whose coherence length is far greater than that of the uncollapsed state.

3. The two-photon light source

The two-photon decay occurs inside a crystal with a $\chi^{(2)}$ nonlinearity (we used a potassium dihydrogen phosphate, or "KDP," crystal) by the decay of a single photon γ_0 produced by an ultraviolet laser (a single-mode argon ion laser at 351 nm) into two red photons γ_1 and γ_2 near 702 nm, in a process known as "parametric fluorescence" or "parametric down-conversion." This process is the reverse of second harmonic generation, in which two red photons combine to form an ultraviolet photon at twice the frequency. Energy and momentum are conserved here:

$$E_0 = E_1 + E_2$$
$$\mathrm{p}_0 = \mathrm{p}_1 + \mathrm{p}_2$$

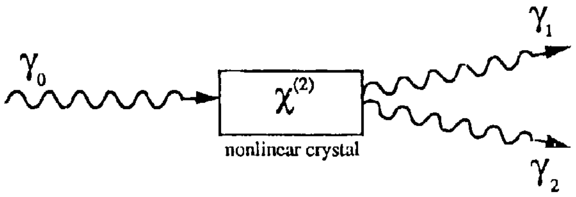


Figure 4. Feynman diagram for parametric down-conversion

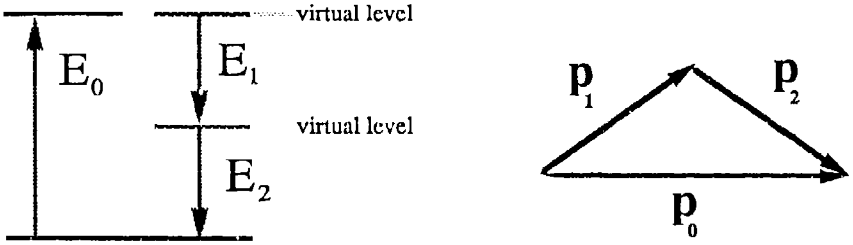


Figure 5. Energy and momentum conservation for parametric down-conversion

The parent photon γ_0 is called the "pump" photon, daughter photon γ_1 is arbitrarily called the "signal" photon, and daughter photon γ_2 the "idler" photon,'for historical reasons. A rainbow of colored cones is produced around an axis defined by the uv laser beam, but *pairs* of photons on opposite sides of the cone are always correlated with each other, e.g., the inner "square" orange photon with outer "square" deep-red photon, etc.

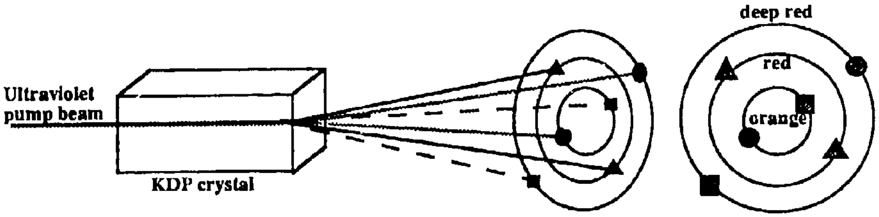


Figure 6. Cones produced in parametric fluorescence. Matching shapes represent conjugate photons, while each ring represents a different color.

The two "conjugate" or "twin" photons are always produced essentially simultaneously in the two-photon decay. They have been observed to be "born" within tens of femtoseconds of each other. They are produced in the entangled state of energy described above. Due to the fact that there are many ways to partition the energy of the parent photon, each daughter photon has a broad spectrum, and hence a narrow wave packet in time. However, due to their entanglement, the *sum* of the two down-converted photons have an extremely sharp energy, since by energy conservation, they must add up to the energy of the parent uv laser photon. Thus the *difference* in their arrival times, and the *sum* of their energies can be simultaneously known to high precision.

4. The "Quantum Eraser"

We have used this nonclassical light source for a "quantum eraser" experiment. The idea of the quantum eraser was recently discussed by Scully, Englert and Walther[13] in connection with the micromaser. Here I present a simpler version of this idea. Consider Young's two-slit experiment from a particle viewpoint.

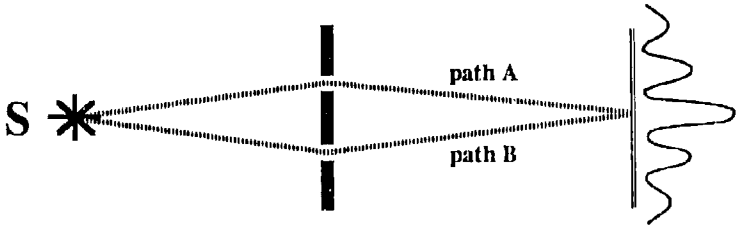


Figure 7. Young's two-slit experiment

The reason one sees interference at the screen is that one cannot know, *even in principle*, which path (A or B) the particle took on its way to the screen. The lack of this "which path" information is fundamental to the observability of interference fringes. However, suppose we placed two circular polarizers of opposite senses, CP1 and CP2, in front of the two slits.

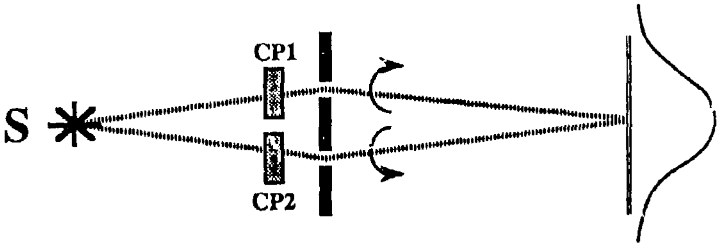


Figure 8. Young's two-slit experiment with circular polarizers CP1 & CP2

The photons which have passed through these circular polarizers are now *labeled* by their polarizations, so that by measuring their helicities, one can know which path the photons took to the screen. Hence we shall call these polarizers "labelers." Since we now have "which path" information, the interference pattern on the screen disappears. (Note that the center-of-mass motion of the particles is in no way disturbed by the insertion of the circular polarizers, so that this scheme is very different from Feynman's[14], where the scattering of a particle near one of the slits uncontrollably disturbs its center-of-mass motion). Now let us "erase" the "which-path" information by the insertion of a linear polarizer LP in front of the screen.

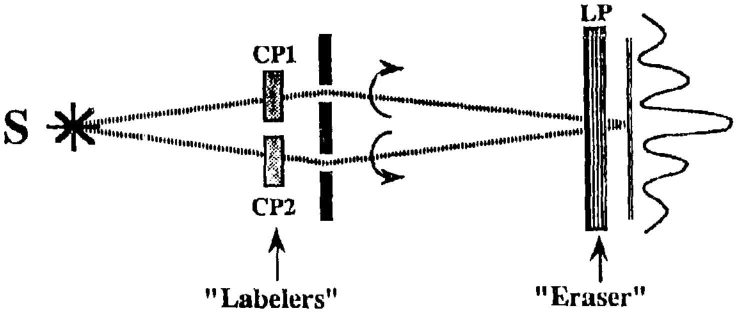


Figure 9. Young's two-slit experiment with circular and linear polarizers

The linear polarizer now *erases* the *handedness* of the photons, which served as their labels. Since "which path" information is now no longer available, the interference pattern is now revived.

This particular version of the "quantum eraser" has a straightforward classical-wave explanation. Hence we decided to use instead the nonclassical two-photon light source described above, in conjunction with the Hong-Ou-Mandel (HOM) two-photon interferometer[15], to demonstrate a "quantum eraser" which had no classical analog. In this interferometer, the two "twin" photons are brought back together by means of mirrors, so that they impinge simultaneously on a 50/50 beam splitter, after which they continue on to the two detectors D1 and D2.

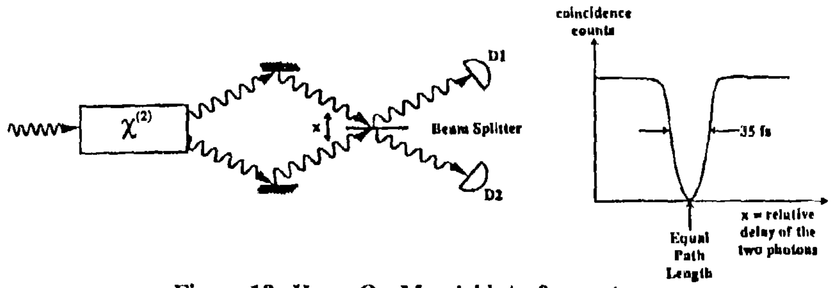


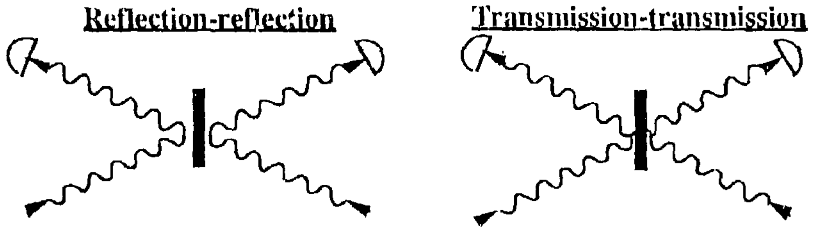
Figure 10. Hong-Ou-Mandel interferometer

The coincidence rate recorded by these detectors is observed to go through a sharp dip as the path length difference between the two photons is scanned by the x-motion of the beam splitter. The width of this dip in our experiments is typically tens of femtoseconds. The narrowness of this width allows very high resolution in time-of-flight comparisons between the two photons. In an experiment we have recently completed, but will not report on here, we have used this high temporal resolution to measure the time it takes a photon to tunnel across a barrier[16, 17].

In order to understand this interference effect, we shall use Feynman's rules for interference: List all possible processes leading to the same final outcome. Here, the possible processes for the two photons at the beam splitter are:

(1) Both photons are transmitted; the outcome: a coincidence "click" of D1 and D2.
(2)&(3) One photon is reflected, the other transmitted; the outcome: no coincidences.
(4) Both photons are reflected; the outcome: a coincidence "click" of D1 and D2.

Next, draw all the *indistinguishable* "paths," or Feynman diagrams, leading to the same final outcome, add their *amplitudes*, and then take the absolute square. Here, coincidence detection processes (1) and (4) are indistinguishable, and thus interfere:



amplitude $= \dfrac{i}{\sqrt{2}} \cdot \dfrac{i}{\sqrt{2}} = -\dfrac{1}{2}$ **amplitude** $= \dfrac{1}{\sqrt{2}} \cdot \dfrac{1}{\sqrt{2}} = +\dfrac{1}{2}$

Figure 11. The two indistinguishable processes leading to coincidences

Because of the phase factor of i in the reflection amplitude for a single photon relative to its transmission amplitude (this is a consequence of time-reversal symmetry applied to the behavior of a single photon at a lossless, symmetric beam splitter), a destructive i/ terference of the "reflection-reflection" and "transmission-transmission" probability amplitudes occurs. Hence the total amplitude for coincidences to occur is (–1/2 + 1/2) = 0: Coincidences never occur! In other words, the two photons always exit the same port of the beam splitter whenever the path length difference is zero, i.e., if the photons arrive at the beam splitter *simultaneously*. However, processes (1) and (4) become distinguishable if the photons arrive at the beam splitter at different times. Hence as the path length difference is scanned, we map out the shape of the photon wave packets. The width of the dip is therefore a measure of the coherence length of the single-photon wave packets.

A schematic of our version of the HOM interferometer is the shown in the following figure:

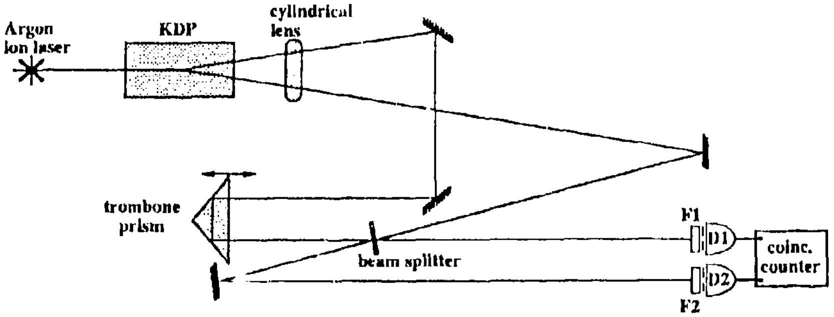


Figure 12. Hong-Ou Mandel interferometer (UCB version)

The mechanism which we used to adjust the path length difference is the "trombone arm," shown in the above figure, consisting of a "trombone prism," which is a right-angle (Porro) prism, mounted on a translation stage, to reflect one of the photons in a trombone-like (or optical delay-line) geometry. This is a technical improvement of the HOM interferometer first implemented by Rarity and Tapster[18]. A typical coincidence "dip" is shown in the next figure:

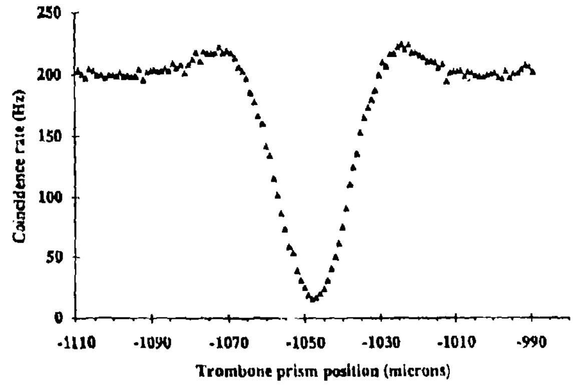


Figure 13. Coincidence rate versus trombone prism position

Now we come to our version of the "quantum eraser" experiment. As in the simpler Young's experiments described earlier, we use polarization as a means of "labeling" the photons, so that we could keep track of "which path" each photon took. The two twin photons emerge from our nonlinear crystal with horizontal linear polarization. Let us add to one arm of the interferometer a "labeler" in the form a half-wave plate (HWP), which can rotate the polarization of the photon to vertical polarization. This clearly enables us in principle to distinguish which path this photon takes, and therefore serves to give us "which path" information. The "erasers" take the form of two polarizers, P1 and P2, oriented at 45 degrees to the vertical, in front of the two detectors.

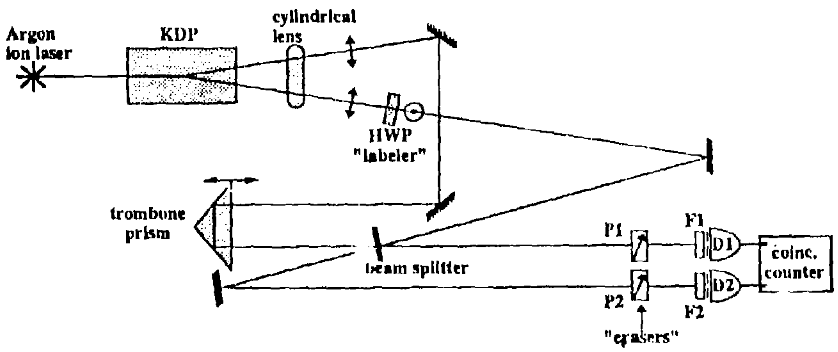


Figure 14. Hong Ou-Mandel interferometer with "labeler" and "erasers"

If the erasers were removed from the above apparatus, the "which path" information, which we could *in principle* obtain from the polarization of the two photons, would destroy the interference pattern. It should be stressed that it is the mere *possibility* of obtaining "which path" information, which destroys the interference pattern; no actual measurements of the polarization of the photons after the beam splitter need be made. In the next figure, we show the disappearance of the coincidence dip, in the absence of P1 and P2, as we rotate the fast axis of the half-wave-plate towards 45° with respect to vertical, at which point the rotation of the polarization of the transmitted photon is 90°, which makes the interfering paths fully distinguishable. (Intermediate orientations of the half-wave plate are also shown).

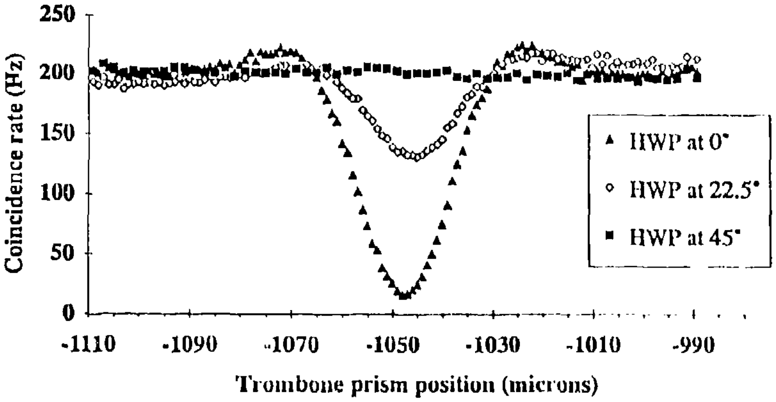


Figure 15. Coincidence rate vs. trombone prism position with "labeler" in setup, but without "erasers"

Now we put in the erasers P1 and P2. By orienting both of them at 45° to the vertical, we can erase the "which path" information, since both horizontally and vertically polarized photons end up polarized at 45° after passing through these polarizers, and we lose the ability to distinguish, even in principle, between the paths taken by the photons. The result is that the interference pattern, i.e., the coincidence dip, is now "revived," as shown by the data represented by the squares in the following figure:

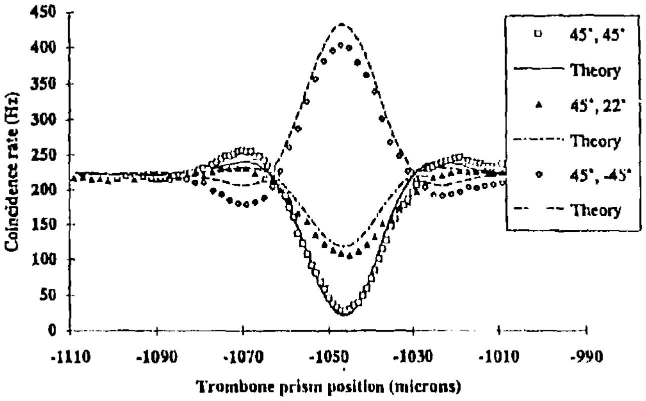


Figure 16. Revival of interference after erasure

Note that the presence of *both* polarizers P1 and P2 is necessary to perform the erasure. The removal of either of them would leave *one* of the photons labeled, carrying enough "which path" information to totally destroy the interference pattern. An interesting feature of this experiment is that one can change the coincidence *dip* into a coincidence *peak* (i.e., an interference minimum into a maximum), by rotating P1 relative to P2 until one is at +45° and the other is at −45°. The data for this orientation (along with those for an intermediate orientation) are represented by the diamonds in the above figure. (We have also checked that the center of the coincidence dip is a sinusoidal function only of the *relative* angle between P1 and P2, which Shih & Alley and Ou & Mandel have already observed in connection with Bell's inequality experiments[19, 20]). Since the resulting interference pattern (dip, peak or something in between) in the end depends on *our* choice of the settings of P1 and P2, we have nicknamed this effect the "quantum editor."

These effects underline the fact that in quantum mechanics, interference only occurs between processes which could not be distinguished from one another even in principle. That is, the final state of the entire system must be considered, including all particles which may have interacted with the interfering particle(s), and both internal and external degrees of freedom. While this fact is a central component of standard quantum mechanics, it is often neglected, though frequently without ill consequences. It is crucial, however, for understanding the other experiments described below.

5. Dispersion-cancellation effect in two-photon interference

As a motivation for the "dispersion-cancellation" experiment, let us return for a moment to the classical problem of propagation in a dispersive medium. We know that the peak of a classical electromagnetic wave packet propagating through a piece of glass will travel at the group velocity, but it is not entirely clear that one can interpret this classical wave packet as if it were the "wavefunction" of the single-photon, and use the Born interpretation for this "wavefunction." If this interpretation were to be correct, then the photon would most likely travel at the group velocity in this medium. However, as Sommerfeld and Brillouin have pointed out[21], at the classical level there already are five kinds of propagation velocities in a dispersive medium: the phase, group, energy, signal and front velocities, all of which differ from the each other in the vicinity of an absorption line, where there is a region of anomalous dispersion. In particular, the group velocity can become "superluminal," i.e., faster than the vacuum speed of light, in these regions. If the photon were to travel at the group velocity in this medium, does it also travel "superluminally"? If not, then at which of these velocities does the photon travel in dispersive media? (We have been studying these questions in the context of photon tunneling times, but shall not discuss them here.)

Motivated by the above questions, we did the following experiment. Let us remove all the polarizers in the "quantum eraser" setup, and return the original HOM interferometer. Now let us insert a piece of glass in the path of one of the photons:

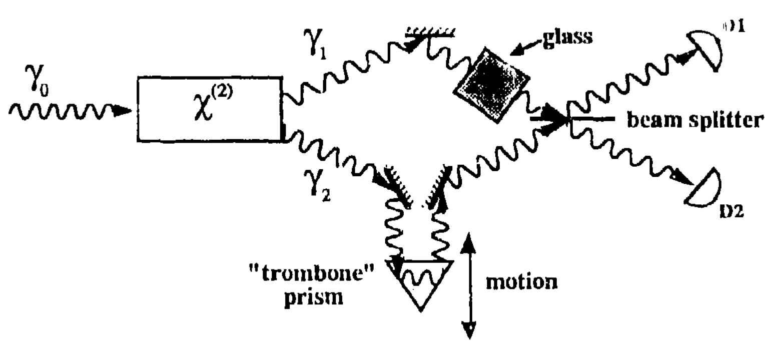


Figure 17. Simplified Hong-Ou-Mandel schematic with glass inserted

This glass slows down the photon which traverses it, and in order to observe the coincidence dip, it is necessary to introduce an equal, compensating delay in the other arm of the interferometer, by adjusting the "trombone" prism. We measured the magnitude of this delay for various samples of glass and were able to determine traversal times on the order of 40 ps, with 4 fs accuracy. In this way, we were able to confirm that single photons travel through glass at the group velocity in transparent spectral regions, an interesting example of particle-wave duality.

Let us consider for a moment the limiting resolution of this measurement technique. The interest of measuring optical delays is greatest for media with dispersion. In dispersive media, however, the broad spectrum required for an ultrafast pulse (or single-photon wave packet) can lead to a great deal of dispersive broadening. One might expect that this broadening of the wave packet would also broaden the coincidence dip in the HOM interferometer, since the width of this dip is a measure of the size of the wave packets which impinge on the beam splitter. For example, one expects a 15 fs wave packet propagating through half an inch of SF11 glass (which we used in our experiment) to broaden to about 60 fs due to the dispersion in this glass. The nature of the broadening is that of a chirp, i.e., the local frequency sweeps from low to high frequency (for normal dispersion, in which redder wavelengths travel faster than bluer wavelengths). Hence the earlier part of the broadened pulse consists of redder wavelengths, and the later part of this pulse consists of bluer wavelengths:

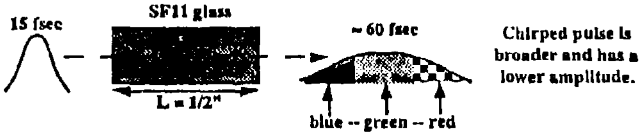


Figure 18. Chirped pulse due to normal dispersion

In our experiment, however, we found that the combination of the time-correlations and energy-correlations exhibited by our entangled photons led to a cancellation of these dispersive effects. While the individual wave packet which travels through the glass does broaden according to quantum mechanics, it is impossible to know whether this photon was reflected or transmitted at the beam splitter (recall Figure 12). This means that when an individual photon arrives at a detector, it is unknowable whether it travelled through the glass, or whether its *conjugate* (with *anticorrelated* frequency) travelled through the glass; due to the chirp, the delay in these two cases is opposite (relative to the peak of the wave packet). An exact cancellation occurs for the (greatly dominant) linear group-velocity dispersion term, and no broadening of the 15 fs interference dip occurs. This is a direct consequence of the nature of the EPR state, in that it relies on the correlations in one observable (energy) to maintain a high degree of accuracy in measuring an *incompatible* observable (time)!

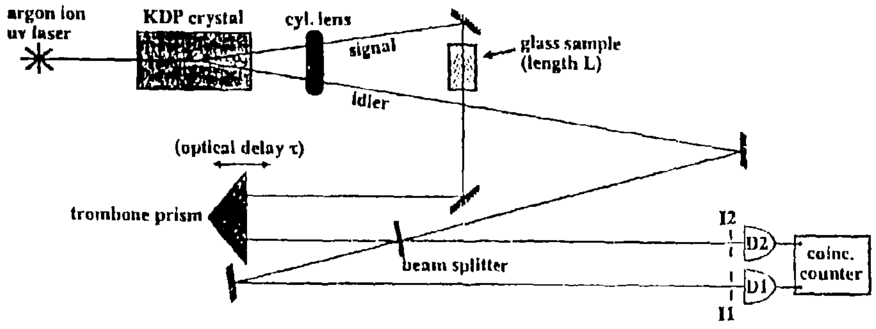


Figure 19. Dispersion cancellation experiment

The apparatus used for this experiment is shown in the above figure. It is essentially the same as that for the quantum eraser, minus all polarizing elements, but plus the glass sample in one of the arms of our version of the HOM interferometer. The resulting coincidence dips with and without the piece of glass are essentially the same shape, as can be seen by comparison of the following data (the dashed curve corresponds to a theoretical 60-fs-wide wave packet)

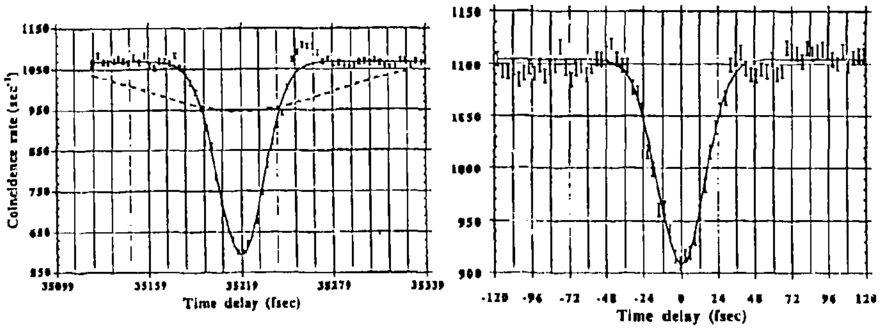


Figure 20. HOM coincidence dips with glass (left trace) and without glass (right trace)

We see that there is indeed very little broadening in the data with the glass compared with that without the glass. Certainly, broadening on the scale of 60 fs (the dashed curve) is ruled out by these data. A detailed theoretical analysis predicted these results, in agreement with the simple argument presented above[22]. This result is important for applications, e.g., in our tunneling-time measurement, since the sharpness of the dip-- and hence the temporal resolution-- is not appreciably degraded by the presence of dispersion in the optical elements of our apparatus or in the sample itself. One lesson learned from these experiments is that the coherence length of the wave packet is not equal to the width of the wave packet, as was also demonstrated by neutron interference experiments.

6. The Franson experiment: Interference between two photons in separated Mach-Zehnder interferometers

Let us begin with the conclusion which we reached from the Franson experiment: A violation of a Bell's inequality for energy and time is implied, thus photons do not necessarily possess a well-defined energy (or color), nor do they possess a well-defined time of emission (or intrinsic age), prior to detection. (As an aside, we note that this contradicts the basic assumption of kinetic theory, viz., that particles carry definite physical properties, such as energy, as well as the basic assumption of this conference, viz., that one can celebrate a birthday at a well-defined time). Our experiment, which is sketched in the following figure, was first proposed by Franson[4].

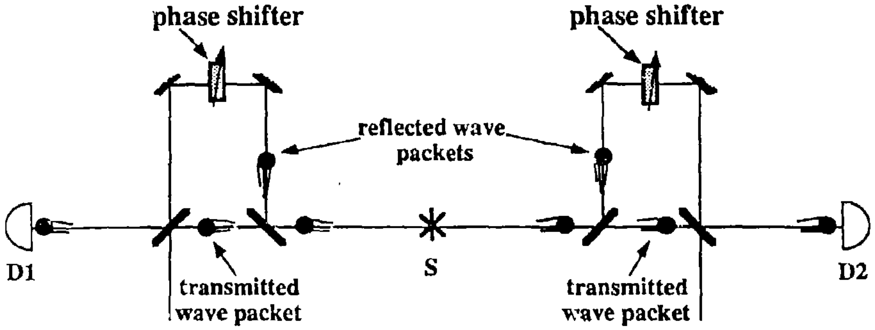


Figure 21. The Franson experiment: The interference of two spatially separated photons in two Mach-Zehnder interferometers

As in the original EPR paper, a source S emits two particles in opposite directions, but the new feature here is that they enter two identical Mach-Zehnder interferometers, in which they are allowed to take either a short path or a long path. These interferometers can have their path length differences adjusted by means of phase shifters inserted into their long paths.

There is no first-order interference of a single photon wave packet with itself inside either interferometer, because the width of the wave packet (which is on the order of tens of femtoseconds in our experiment) is much too small to permit any overlap of the transmitted and reflected portions of the wave packet at the final beam splitter. However, there is a second-order (i.e. two-photon) interference observable in coincidence detection at detectors D1 and D2.

Again, we shall use Feynman's rules for interference to calculate the probability of coincidence detection. The indistinguishable processes here are (1) the "short-short" and (2) the "long-long" processes, (where in (1), both photons take the short paths of their respective interferometers, and in (2) they both take the long paths). The distinguishable processes are (3) the "short-long" and (4) the "long-short" processes, since the "clicks" of D1 and D2 are not simultaneous in these two processes. In principle and also in practice, we are able to reject these distinguishable "clicks" by using sufficiently large path length differences in the two interferometers, and a sufficiently narrow coincidence timing window in our electronics. We are thus left with the two indistinguishable processes (1) and (2) only, for which we must first add the probability *amplitudes*, and then take the absolute square. Hence the probability of a given coincidence detection is given by the expression

$$P_c \propto \left| 1 \cdot 1 + e^{i\phi_1} \cdot e^{i\phi_2} \right|^2$$

where the first term inside the absolute value corresponds to the "short-short" process, and the second term to the "long-long" process. (The beam splitters are assumed to be 50/50 throughout.) Here the phase ϕ_1 (ϕ_2) represents the total phase difference between the long and short arms of the left (right) interferometer. Simplifying this expression, we get

$$P_c \propto \left[1 + \cos(\phi_1 + \phi_2)\right] .$$

Note that this implies a fringe visibility of 100% (i.e., perfect zeros at the minima in coincidence detections). Bell's inequality for this experiment implies (when certain reasonable auxiliary assumptions are made) that sinusoidal fringes can have at most 70.7% ($=1/\sqrt{2}$) visibility.

Our apparatus is sketched in the following figure. In the second figure, we present our data:

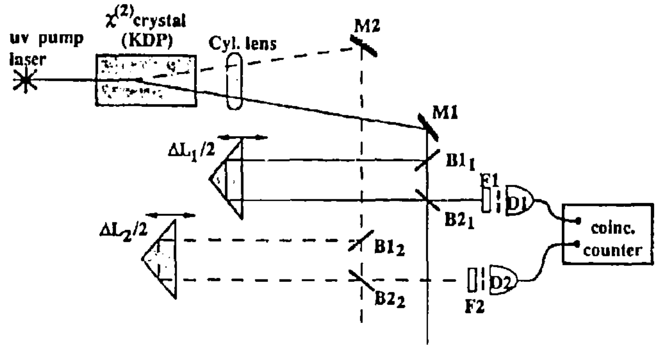


Figure 22. Apparatus used at Berkeley to perform the Franson experiment

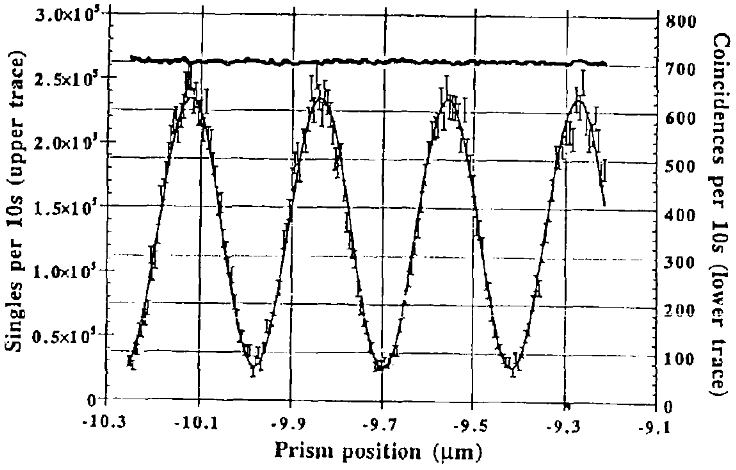


Figure 23. Interference fringes of our Franson experiment

By analysis of these data, we concluded that Bell's limiting 70.7% visibility is exceeded by 16 standard deviations.

The meaning of the maximum is that the two spatially separated photons always behave in a *correlated* fashion at the final beam splitter, i.e., if one is transmitted, then the other is also transmitted; and if one is reflected, then the other is also reflected. The meaning of the minimum is that the two photon "twins" always behave in an *anticorrelated* fashion at the final splitter, i.e., if one is transmitted, then the other is reflected, and vice versa. The behavior of the "twins" depends on the settings which *we* choose for the space-like separated phase shifters (which we could in principle set even *after* the photons had entered their separate interferometers[23]). Also, it should be emphasized that the fact that these interference fringes were observed means that one does not know, even in principle, the actual age of the "twins" upon their arrival (i.e. detection) for otherwise the "long-long" and "short-short" processes would become *distinguishable*, and the interference pattern would disappear.

As a final remark, other papers at this conference addressed the question of whether pure quantum states can evolve into mixed states in black hole evaporation. A closely related question is whether *mass* is a local, realistic property of a black hole. Let us consider the following photon pair-creation process arising from a vacuum fluctuation at the event horizon of a black hole:

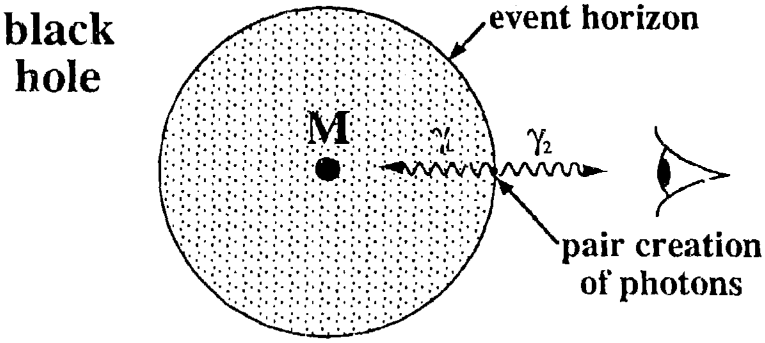


Figure 24. Photon hair creation at the event horizon of a black hole

The left-going member of the pair falls into the black hole, whereas the right-going member escapes to infinity. Since energy is conserved in this system, the mass of the black hole is entangled with that of the photon which escapes to infinity, and the entire system is in an entangled state. In light of the violation of Bell's inequality in our experiment, it may be forbidden to ascribe any well-defined mass to the black hole until this right going photon is detected, i.e., it may be incorrect to think of the mass of the black hole as a local, realistic quantity until it is observed. Einstein and Bohr had a similar discussion (though not in the context of black holes) at the 1930 Solvay Conference[24].

7. Acknowledgments

We thank Bruce E. Johnson for his help. This work is supported by ONR grant number N00014-90-J-1259.

8. References

1. A. Einstein, B. Podolsky and N. Rosen, *Phys. Rev.* **47**, 777 (1935).
2. P.G. Kwiat, A.M. Steinberg and R.Y. Chiao, *Phys. Rev.* **A45**, 7729 (1992).
3. A.M. Steinberg, P.G. Kwiat and R.Y. Chiao, *Phys. Rev. Lett.* **68**, 2421 (1992).
4. J.D. Franson, *Phys. Rev. Lett.* **62**, 2205 (1989).
5. P. G. Kwiat, A. M. Steinberg and R. Y. Chiao, *Phys. Rev.* **A 47**, R2472 (1993).
6. D. Bohm, in *Quantum Theory and Measurement,* J.A. Wheeler and W.H. Zurek ed., (Princeton Univ. Press, Princeton, 1983), p. 356.
7. J.F. Clauser and A. Shimony, *Rep. Prog. Phys.* **41**, 1881 (1978).
8. A. Aspect, J. Dalibard and G. Roger, *Phys. Rev. Lett.* **49**, 1804 (1982).
9. J.S. Bell, *Physics* **1**, 195 (1964).
10. J.G. Rarity and P.R. Tapster, *Phys. Rev. Lett.* **64**, 2495 (1990).
11. E. Schrödinger, in *Quantum Theory and Measurement,* J.A. Wheeler and W.H. Zurek ed., (Princeton Univ. Press, Princeton, 1983), p. 152.
12. R.Y. Chiao, P.G. Kwiat and A.M. Steinberg, *Proc. Workshop on Squeezed States and Uncertainty Relations* (NASA Conference Publication 3135, 1991).
13. M.O. Scully, B.-G. Englert and H. Walther, *Nature* **351**, 111 (1991).
14. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, 1965).
15. C.K. Hong, Z.Y. Ou and L. Mandel, *Phys. Rev. Lett.* **59**, 2044 (1987).
16. A.M. Steinberg, P.G. Kwiat and R.Y. Chiao, submitted to *Phys. Rev. Lett.* (1993).
17. A.M. Steinberg, P.G. Kwiat and R.Y. Chiao, submitted to *Proc. XXVIIIe Rencontre de Moriond* (France, 1993).
18. J.G. Rarity and P.R. Tapster, *Phys. Rev.* **A 41**, 5139 (1990).
19. Y.H. Shih and C.O. Alley, *Phys. Rev. Lett.* **61**, 2921 (1988).
20. Z.Y. Ou and L. Mandel, *Phys. Rev. Lett.* **61**, 50 (1988).
21. L. Brillouin, *Wave Propagation and Group Velocity* (Academic, New York, 1960).
22. A.M. Steinberg, P.G. Kwiat and R.Y. Chiao, *Phys Rev.* **A45**, 6659 (1992).
23. J.A. Wheeler, in *Quantum Theory and Measurement,* J.A Wheeler and W.H. Zurek ed., (Princeton Univ. Press, Princeton, 1983), p. 182.
24. N. Bohr, in *Quantum Theory and Measurement,* J.A. Wheeler and W.H. Zurek ed., (Princeton Univ. Press, Princeton, 1983), p. 32.

EINSTEIN-PODOLSKY-ROSEN CORRELATIONS IN HIGHER DIMENSIONS

A. ZEILINGER, M. ZUKOWSKI*
Institut für Experimentalphysik, Universität Innsbruck
A-6020 Innsbruck, Austria

M.A. HORNE
Stonehill College, North Easton, MA 02357, USA

H.J. BERNSTEIN
Hampshire College, Amherst, MA 01002, USA

D.M. GREENBERGER
City College of the City University of New York, New York, NY 10031, USA

ABSTRACT

Using multiport beam splitters it will be possible to study Einstein-Podolsky-Rosen correlations in higher dimensional Hilbert space. As an explicit example we present the design and theory of a tritter, which is a multiport beam splitter with three input ports and three output ports, such that any amplitude incident at one input port is distributed equally over the output ports. We will then show the results for a two-photon, two-tritter experiment, where novel Einstein-Podolsky-Rosen correlations occur.

1. Introduction

All experimental work concerning the Einstein-Podolsky-Rosen Paradox[1] and Bell's theorem[2] thus far is restricted to two-particle (in most cases two-photon) entangled states where the correlations can effectively be described by restricting the analysis to a Hilbert space of dimension 2 for each particle. These states can be two polarization states as proposed initially by Bohm[3] and first employed in an experiment by Freedman and Clauser[4], they can be two momentum eigenstates as in the experiment proposed by Horne and Zeilinger[5] and performed first by Rarity and Tapster[6], or, they can be two states which took beam paths of markedly different length on their way from the source to the detector as proposed by Franson[7]. This latter experiment has now been performed by various groups[8], the most conclusive experiment which showed a striking violation of a Bell-type inequality is due to Kwiat, Steinberg and Chiao[9].

There are two obvious routes for generalization. One is to consider more than two particles, the other is to analyze the case of more than two states available to each particle. The generalization to more than two particles has led to some new insight into the difference

[*] Permanent address: *Institute of Theoretical Physics and Astrophysics, University of Gdansk, PL-80952 Gdansk, Poland*

between quantum mechanics and local realistic theories[10]. But, due to the unavailability of coherent multi-particle sources this has not as yet resulted in an experiment.

In the present paper we would like to focus our analysis on another generalization. This is the case where each particle has more than two states available. The correlations are then defined in Hilbert spaces of higher dimension[11]. It is obvious that a possible route to generalizing EPR correlations to systems of higher dimension would be to investigate spin correlations between particles with spin-1 or higher (with the obvious and notable exception of the photon or other massless particles which have only 2 polarisation states.) Again, since at present there exist no sources for correlated particles of higher spin, such investigations based on spin correlations are purely theoretical to date[12].

This paper shows how to obtain such EPR correlations in more than two dimensions in real experiments. Such experiments are based on both the availability of parametric down-conversion as a source for highly correlated two-photon states[13] and on the use of multi-port devices[14]. Finally, we present some theoretical predictions for the novel correlations expected.

2. The Beam Splitter as a Four-Port Device

The beam splitter is a central element of many experiments in quantum optics. A general beam splitter has two input ports and two output ports (Fig. 1). Formally it may be described by a unitary operator in a two-dimensional Hilbert space. We should note here that for the present paper we deliberately adopt an explicit Hilbert space formalism because it is equally well suited for describing a beam splitter operating for any type of particle, be it electrons, photons, atoms or neutrons, to name just those types of radiation for which quantum interference experiments with beam splitters have been performed so far.

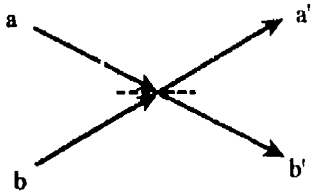


Fig. 1: A general beam splitter has two input ports and two output ports.

The general beam splitter pure input state is a superposition

$$|\psi> = \psi_a|a> + \psi_b|b> \tag{1}$$

where $|a>$ and $|b>$ describe a particle in beam a or b (see Fig. 1) respectively. We assume the normalization $\psi_a\ \psi_a{}^* + \psi_b\ \psi_b{}^* = 1$. Likewise the general output state is the superposition

$$|\psi'> = \psi'_a|a'> + \psi'_b|b'> \tag{2}$$

in obvious notation. Input and output states may equally well be written in matrix notation as

$$\psi=\begin{pmatrix}\psi_a\\ \psi_b\end{pmatrix},\quad \psi'=\begin{pmatrix}\psi'_a\\ \psi'_b\end{pmatrix}. \tag{3}$$

The general beam splitter operator U then couples ψ' to ψ, $\psi'=U\psi$ with $U^{+}U=I$.

We restrict ourselves now to 50-50 beam splitters. This means that a particle incident at any of the two input ports of a symmetric beam splitter has the same probability $p=1/2$ to be found in any of the two output ports. It is well known that such a beam splitter is defined only up to arbitrary phase factors in the input and output ports[15].

Two possible 50-50 beam splitter operators are for example

$$U_t=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} \quad \text{or} \quad U_s=\frac{1}{\sqrt{2}}\begin{pmatrix}i & 1\\ 1 & i\end{pmatrix} \tag{4}$$

where U_t represents a time-symme'ric beam splitter and U_s represents a spatially symmetric one. The two beam splitters can be converted into each other using π phase shifts in one input and one output port, i.e.

$$U_t=\begin{pmatrix}-i & 0\\ 0 & 1\end{pmatrix}U_s\begin{pmatrix}1 & 0\\ 0 & i\end{pmatrix}. \tag{5}$$

The two beam splitter operators imply different transition rules for incident beams. These are

$$\begin{aligned} &|a\rangle\Rightarrow\frac{1}{\sqrt{2}}\{|a'\rangle+|b'\rangle\} \qquad |b\rangle\Rightarrow\frac{1}{\sqrt{2}}\{|a'\rangle-|b'\rangle\} \text{ for } U_t,\\ &|a\rangle\Rightarrow\frac{1}{\sqrt{2}}\{i|a'\rangle+|b'\rangle\} \qquad |b\rangle\Rightarrow\frac{1}{\sqrt{2}}\{|a'\rangle+i|b'\rangle\} \text{ for } U_s. \end{aligned} \tag{6}$$

The first beam splitter implies no phase change upon reflection from one side while reflection from the other side implies a phase change of π. The second beam splitter operator implies that both reflected beams acquire a phase shift of π/2 upon reflection.

We note here that beam splitters are just special cases of 4-port devices. Another example of a 4-port device would be a Mach-Zehnder interferometer.

3. Two-Particle Two-State Systems

Using these rules it is now easily possible to calculate the results of a two-particle two-state EPR-Bell experiment as shown in Fig. 2 A source emits two particles in the state

$$|\psi\rangle=\frac{1}{\sqrt{2}}\{|a\rangle|c\rangle+|b\rangle|d\rangle\}. \tag{7}$$

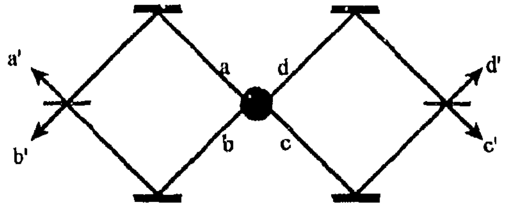


Fig. 2: Principle of a two-particle, two-state EPR-Bell experiment using beam splitters.

Here and below the first ket in a product always refers to particle 1 and the second to particle 2. Also, e.g., $|a\rangle|c\rangle$ implies the tensor product $|a\rangle\otimes|c\rangle$ etc. The beams a, b, c, d may then be subject to the phase shifts α, β, χ, δ such that the state becomes

$$|\psi\rangle = \frac{1}{\sqrt{2}} e^{i(\alpha+\gamma)}\left\{|a\rangle|c\rangle + e^{i\chi}|b\rangle|d\rangle\right\} \qquad . \qquad (8)$$

with $\chi = \beta + \delta - \alpha - \gamma$. Applying now the beam splitter rules (6) and, analogously,

$$|c\rangle \Rightarrow \frac{1}{\sqrt{2}}\left\{|c'\rangle + |d'\rangle\right\} \qquad |d\rangle \Rightarrow \frac{1}{\sqrt{2}}\left\{|c'\rangle - |d'\rangle\right\} \qquad (9)$$

one obtains for the joint probabilities for two detectors to register the particles in coincidence

$$\begin{aligned} p(a',c') = p(b',d') &= \frac{1}{2}\cos^2(\chi/2) \\ p(a',d') = p(b',c') &= \frac{1}{2}\sin^2(\chi/2). \end{aligned} \qquad (10)$$

Thus, perfect correlations arise for

$$\chi = n\pi. \qquad (11)$$

For odd n detector a' fires in coincidence with detector d' and detector b' fires in coincidence with detector c' while for even n the coincidences are $a' - c'$ and $b' - d'$. These two different types of coincidences are represented in Fig. 3. In other words, for these parameter settings the path taken by a particle after its beam splitter is an Einstein-Podolsky-Rosen element of reality, i.e. firing of any one individual detector for one particle allows one to predict with certainty which detector will register the other particle.

These perfect correlations can be characterized via a value-assignment procedure introduced by Bell. The possible results obtained on either side are named A and B, and they are assigned the values ± 1. It then follows that the two cases of perfect correlation are signified by $AB = +1$ and $AB = -1$ respectively. We call these values Bell numbers. We notice that one of the beam splitter operator representations (U_t) just contains Bell numbers (+1 and -1 for the two dimensional case). It will be seen later that for multiports the generalization of

Bell's value assignment procedure is quite interesting. Furthermore, in any dimension there are always multiports whose unitary representation contains only Bell numbers.

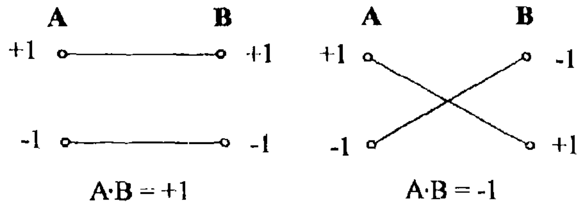


Fig. 3: Possible perfect correlations for the case of an experiment as shown in Fig. 2. The results A, B on either side can be +1 or -1, depending on which detector in which outgoing beam registers a particle. The perfect correlations can be signified by either $A \cdot B = +1$ or $A \cdot B = -1$.

We should mention that the results of this section are basically known. They were repeated here in order to prepare the reader for the less familiar situations in the following sections. An experimentally available source which prepares the two particles in the entangled state of Eq. (7) is a non-linear crystal where through the process of spontaneous parametric down-conversion an incident photon may split into 2 photons of lower energy.

4. The Tritter as an Example of a Multiport Device

In this section we first introduce the general concept of multiports and then we give some explicit examples. A general multiport has L input ports and M output ports* and is called N-port $(N = M + L)$. For simplicity we restrict our considerations to symmetric N-ports which are defined as having an equal number of input ports and output ports $(L = M = N/2)$ and, furthermore, which operate such that a single particle incident on any individual input port has equal probability (i.e. $p = 1/M = 2/N$) to be found in any specific output port. This is the generalization of the generic beam splitter discussed in section 2 above. We propose to call symmetric multiports "Critters" and specifically a critter with $L = M = 3$ is called a Tritter, one with $L = M = 4$ is a Quitter[16] etc.

Lossless symmetric multiports (critters) can be represented by unitary operators in an M-dimensional Hilbert space. Again, as was the case for the conventional beam splitter, there are many physically possible critters, but, as opposed to the beam splitter case, it is not always possible to transform all types of a specific critter (i.e. symmetric N-port with a given N) into each other by merely supplying external phase shifters or relabelling output ports[17].

Let us consider explicitly the tritter. The general input and output states are (Fig. 4)

$$\begin{aligned} |\psi\rangle &= \psi_a|a\rangle + \psi_b|b\rangle + \psi_c|c\rangle \\ |\psi'\rangle &= \psi'_a|a'\rangle + \psi'_b|b'\rangle + \psi'_c|c'\rangle. \end{aligned} \qquad (12)$$

or, in matrix notation,

*In general some physical ports can work both as input and output ports (viz. the Michelson interferometer).

$$\psi = \begin{pmatrix} \psi_a \\ \psi_b \\ \psi_c \end{pmatrix} \quad \text{and} \quad \psi' = \begin{pmatrix} \psi'_a \\ \psi'_b \\ \psi'_c \end{pmatrix}. \tag{13}$$

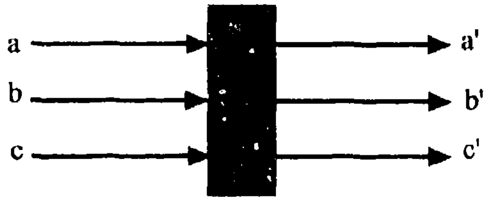


Fig. 4: A generic tritter is devised with three input ports and three output ports such that an amplitude incident on any one of the input port excites any of the output ports equally.

Again, a unitary operator couples the output state to the input state,

$$\psi' = U\psi. \tag{14}$$

This unitary operator can now be represented by a 3 x 3 matrix where the modulus of each matrix element is $1/\sqrt{3}$. Here again and for all critters it is possible to absorb any phase factors of the first row into phases of the input beams and to absorb any phase factors of the first column into phases of the output beams. Such a representation of a multiport only contains "1" in both the first column and the first row. We will call such a representation canonical. Thus, the general tritter operator can be written as

$$U = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \varphi & \varphi^* \\ 1 & \varphi^* & \varphi \end{pmatrix} \tag{15}$$

with $|\varphi| = 1$ and $\varphi + \varphi^* = -1$. The only *two* possible choices for φ are $\varphi = \alpha$ and $\varphi = \alpha^2$ with $\alpha = e^{2\pi i/3}$.

Thus the tritter operator has two canonical representations, either

$$U_T = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{pmatrix} \quad \text{or} \quad U'_T = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha^2 & \alpha \\ 1 & \alpha & \alpha^2 \end{pmatrix}. \tag{16}$$

The transition rules for incident beams therefore are

$$
\begin{aligned}
|a\rangle &\Rightarrow \frac{1}{\sqrt{3}}\{a'\rangle + b'\rangle + c'\rangle\} \\
|b\rangle &\Rightarrow \frac{1}{\sqrt{3}}\{|a'\rangle + \alpha|b'\rangle + \alpha^2|c'\rangle\} \\
|c\rangle &\Rightarrow \frac{1}{\sqrt{3}}\{|a'\rangle + \alpha^2|b'\rangle + \alpha|c'\rangle\}
\end{aligned}
\tag{17}
$$

for the tritter rule U_T. For U_T' the roles of α and α^2 are just interchanged. Note also that $U_T^{-1} = U_T'$ and that the two different types of tritter can bo converted into each other by an odd number of permutations of rows or columns, e.g.

$$
U_T' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} U_T. \tag{18}
$$

These results imply that sequential arrangement of tritters does not lead to new nontrivial tritters. In other words, given some tritter one can obtain any tritter by changing external phases and by a permutation of input and/or output ports, which may simply be achieved for example by flipping two output ports. Physically, there are many different possibilities of realising a tritter. A specific type with parallel input beams and parallel output beams is shown in Fig. 5. One can easily see that a tritter has more adjustable parameters than a beam splitter. These are the reflectivities of the partially reflecting mirrors and the nontrivial phase in the internal loop of the tritter.

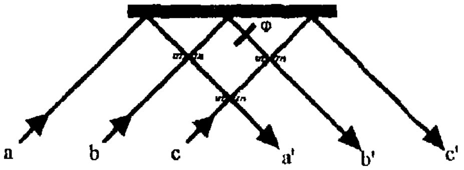


Fig. 5: Possible realization of a tritter using partially reflecting mirrors and a nontrivial internal phase $\phi \neq 0, \pi$.

Turning to higher multiports the number of experimentally adjustable nontrivial parameters grows quadratically with the number of ports. One of the most interesting results for higher multiports is the existence of distinct classes which cannot be transformed into each other by just changing external phases and by permutation of input and/or output ports. We leave a detailed discussion to a forthcoming paper.

5. Two-Particle Three-State Systems

It is evident that with multiports a large number of novel experiments in quantum optics become possible. Because of the availability of down-conversion photon sources, we only discuss here the case where a two-particle source is employed. Assume that such a source emits two particles in the state

$$|\psi> = \frac{1}{\sqrt{3}}\{|a\rangle|d\rangle + |b\rangle|e\rangle + |c\rangle|f\rangle\}. \tag{19}$$

Again, the first ket in a product state refers to particle 1, and the second to particle 2. The beams a,b,c,d,e,f are subject to the phase shifts $\alpha,\beta,\gamma,\delta,\varepsilon,\zeta$, respectively, and thus the state evolves into

$$|\psi\rangle = \frac{1}{\sqrt{3}}e^{i(\alpha+\delta)}\{|a\rangle|d\rangle + e^{i\varphi}|b\rangle|e\rangle + e^{i\chi}|c\rangle|f\rangle\} \tag{20}$$

with $\chi = \beta + \varepsilon - \alpha - \delta$ and $\varphi = \gamma + \zeta - \alpha - \delta$.

Suppose now that the three beams excited by particle 1 are fed into a tritter and likewise the three beams excited by particle 2 are fed into another tritter (Fig. 6). Clearly the final state is then obtained by applying the appropriate tritter operator Eq. (16) to state (20). Instead of writing down the final state explicitly, we focus on the count rates and on the correlations to be expected.

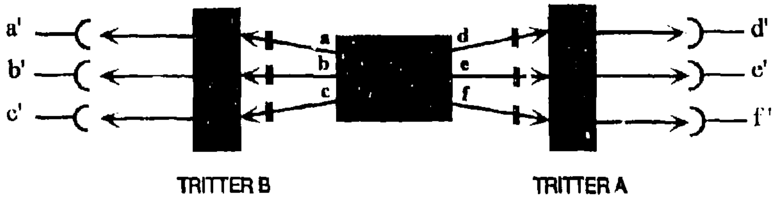


Fig. 6: Principle of a two-tritter, two-photon EPR experiment. In a practical realization the source can be parametric down-conversion.

The unconditional probability to detect a particle in any of the detectors is a constant, e.g. $p(a') = 1/3$. The independence of any of the phases inserted between source and tritters is a consequence of the initial entanglement. Certainly this does not hold anymore for the various joint probabilities of detecting a particle in a given detector on one side together with detecting the other particle on the other side. These joint probabilities are:

$$\begin{gathered}
p(a',d') = p(b',f') = p(c',e') = \frac{1}{27}[3 + 2\cos\chi + 2\cos\varphi + 2\cos(\varphi - \chi)], \\
p(a',e') = p(b',d') = p(b',f') = \frac{1}{27}[3 + 2\cos\chi' + 2\cos\varphi' + 2\cos(\varphi' - \chi')] \\
\text{with } \chi' = \chi + 2\pi/3,\ \varphi' = \varphi - 2\pi/3 \\
p(a',f') = p(b',e') = p(c',d') = \frac{1}{27}[3 + 2\cos\chi'' + 2\cos\varphi'' + 2\cos(\varphi'' - \chi'')] \\
\text{with } \chi'' = \chi - 2\pi/3,\ \varphi'' = \varphi + 2\pi/3
\end{gathered} \tag{21}$$

and where, e.g., $p(a',e')$ is the probability to simultaneously detect a particle in detector a' and a particle in detector e'.

The joint probabilities of Eqs. (21) have a number of remarkable features. It is easy to show that all these probabilities are nonnegative and their maximum value is 1/3. This may be understood by analyzing for example the case where the first equation attains its maximum value which occurs when $\chi, \varphi = 2n\pi$. Then $p(a',d') = p(b',f') = p(c',e') = 1/3$ and all other joint probabilities vanish. This implies that if the phases in the two-tritter two-particle interferometer are set to these values then perfect correlations arise, and thus Einstein-Podolsky-Rosen elements of reality may be introduced. Explicitly, if, say, detector a' fires and the phases are set to the parameters just mentioned we can predict with certainty that the other particle will be registered by detector d'. Likewise, if particle 1 is registered by detector $b'(c')$, particle 2 will be registered by detector $f'(e')$. Thus, while it is always maximally uncertain which detector will register either of the particles, it is known with certainty which detector will register the second particle once the first particle has been observed, as long as the phases are set according to the above condition.

Another set of similar perfect correlations arises if the phases are set such that $\chi', \varphi' = 2n\pi$. Then the joint probabilities are $p(a',e') = p(b',d') = p(e',f') = 1/3$ with all others being zero. Here again perfect correlations occur but now between different detectors than before. Finally, a third set of perfect correlations arises for $\chi'', \varphi'' = 2n\pi$, then $p(a',f') = p(b',e') = p(c',d') = 1/3$ with all other joint probabilities vanishing. Fig. 7 shows these three possible ways of perfect correlations. Note that of the six possible one-to-one combinations between detectors on either side only three combinations are realized for perfect correlations. Here we should note the fact that these types of perfect correlations arise whenever we use the same tritter on each side (either the one represented by U_T or the one represented by U_T^*). In case we chose to use different types of tritters on the two sides, the other three types of perfect correlations occur, with the original three now being excluded.

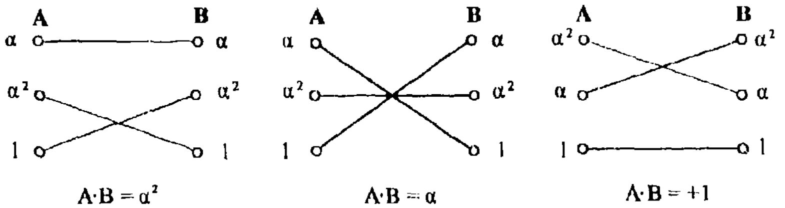


Fig. 7: Perfect correlations occurring in an experiment of the type of Fig. 6. The results on either side are best characterized by assigning them the value $A, B = \alpha, \alpha^2, 1$, where $\alpha = e^{2\pi i/3}$. The three cases of perfect correlations occurring are then signified by $A \cdot B = \alpha, \alpha^2, 1$.

The three types of perfect three-state correlations may be signified in the same way by value assignment as it was done originally by Bell for two-state correlations. One might be tempted to assign the values +1,0,-1 to the three possible outcomes on each side. Such a procedure does not succeed because, when calculating the product AB, if A is again the result for one particle and B the result for the other, appearance of a "0"-result always leads to $AB = 0$ independent of which type of correlation occurs and thus information is lost. A rather elegant procedure of value assignment is to choose α, α^2, α^3 (with $\alpha = e^{2\pi i/3}$) for the three possible outcomes on either side. It then follows that the three cases of perfect correlations are signified by $AB = \alpha, \alpha^2, 1$ (see Fig. 7). These numbers are now the Bell numbers for a three-dimensional Hilbert space.

In general, for the case of correlations between two particles, where each one is defined in an M-dimensional Hilbert space, at most M cases of perfect correlations (where EPR-elements of reality can be introduced) occur with a given set of multiports. It is thus natural to generalize the procedure just given by assigning the values $A,B = e^{2\pi i n/M}$ $n = 1,2,\ldots M$ to the results in order to signify the cases of perfect correlations by $AB = e^{2\pi i n'/M}$. As we will show in a forthcoming paper there is always at least one case of a specific multiport for any M where this procedure succeeds. But, we should point out, for $M\rangle 3$ these are also cases where this procedure fails. Obviously the case $M = 2$ as analyzed originally by Bell is just the most simple nontrivial case. This is the reason why we propose to call these general numbers used in value assignment Bell-numbers.

Concluding this section we note that besides introducing EPR elements of reality the way just given, one can also apply a generalized Bell inequality to the two-tritter correlations[18] thus providing the first feasible test for Bell's theorem for pairs of spins higher than 1/2 via an optical analog.

6. Concluding Comments

In general, an experiment using multiports which are fed the two correlated photons created in the process of parametric down-conversion provides a generalization of EPR correlations to Hilbert spaces of higher dimensions. These correlations are fully analogous to those between two particles with higher spin. Thus they are expected to give new interesting results going beyond those realizable in spin correlations between two spin-1/2 particles or two photons. A specific example are those correlations which are necessary to establish the Bell-Kochen-Specker paradox[19]. Using two correlated particles each defined in a higher-dimensional Hilbert space it is possible to establish the results for each individual measurement utilized in the Kochen-Specker argument as Einstein-Podolsky-Rosen elements of reality[20]. It is evident that using multiports together with a down-conversion photon source can provide immediate experimental realization of such correlations.

This work was supported by the Austrian Science Foundation FWF, project No. S6502 (Schwerpunkt Quantenoptik), and the US National Science Foundation, grant No. PHY92-13964.

7. References

1. A. Einstein, B. Podolsky, and N. Rosen, *Phys. Rev.* **47** (1935) 771.
2. J.S.Bell, *Physics* (NY) **1** (1964) 195.
3. D. Bohm, *Quantum Theory*, (Prentice-Hall, Englewood Cliffs, 1951).
4. S.J. Freedman and J.F.Clauser, *Phys. Rev. Lett.* **28** (1972) 938.

5. M.A. Horne and A. Zeilinger, in *Proc. of the Symposium on the Foundations of Modern Physics*, eds. P. Lahti and P. Mittelstaedt (World Scientific, 1985); M. Zukowski and J. Pykacz, *Phys. Lett.* **A127** (1988) 1; M.A. Horne, A. Shimony, and A. Zeilinger, *Phys Rev. Lett.* **62** (1989) 2209.
6. J.G. Rarity and P.R. Tapster, *Phys. Rev. Lett.* **64** (1990) 2495.
7. J.D. Franson, *Phys. Rev. Lett.* **62** (1989) 2205.
8. P.G. Kwiat, W.A. Vareka, C.K. Hong, H. Nathel, and R.Y. Chiao; *Phys. Rev.* **A41** (1990) 2910; Z.Y. Ou, X.Y. Zou, L.J. Wang, and L. Mandel, *Phys. Rev. Lett.* **64** (1990) 321; J.G. Rarity, P.R. Tapster, E. Jakeman, T. Larchuk, R.A. Campos, M.C. Teich, and B.E.A. Saleh, *Phys. Rev. Lett.* **65** (1993) 1348; J. Brendel, E. Mohler, and W. Martiensen, *Phys. Rev. Lett.* **66** (1991) 1142; *Europhys. Lett.* **20** (1992) 575.
9. P.G. Kwiat, A.M. Steinberg, and R.Y. Chiao, *Phys. Rev.* **A47** (1993) R 4272.
10. D.M. Greenberger, M. Horne, and A. Zeilinger, in *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, ed. M. Kafatos (Kluwer Academic, Dordrecht, 1989), p. 73; D.M. Greenberger, M.A. Horne, A. Shimony, and A. Zeilinger, *Am. J. Phys.* **58** (1990) 1131; N.D. Mermin, *Phys. Rev. Lett.* **65** (1990) 1838.
11. L. Hardy (*Phys. Rev. Lett.* **68** (1992) 2981) has recently made a proposal employing two particles, each one defined in more than two states with the attempt at arriving at a GHZ-type contradiction between local realism and quantum mechanics. In this case and in a related proposal (H.J. Bernstein, D.M. Greenberger, M.A. Horne, and A. Zeilinger, *Phys. Rev.* A47 (1993) 78) that goal is only achieved with some additional assumptions.
12. N.D. Mermin, *Phys. Rev.* **D22** (1980) 356; A. Garg and N.D. Mermin, *Phys. Rev Lett.* **49** (1982) 901; N.D. Mermin and G.M. Schwarz, *Found. Phys.* **12** (1982) 101. M. Ardehali, *Phys. Rev.* **D44** (1991) 33[illegible] G.S. Agarwal, *Phys. Rev.* **A47** (1993) 4608.
13. D.C. Burnham and D.L. Weinberg, *Phys. Rev. Lett.* **25** (1970) 84.
14. A. Zeilinger, H.J. Bernstein, D.M. Greenberger, M.A. Horne, and M. Zukowski, in *Quantum Control and Measurement*, eds. H. Ezawa and Y. Murayama (Elsevier, 1993).
15. A. Zeilinger, *Am. J. Phys.* **49** (1981) 88; S. Prasad, M.O. Scully, and W. Martienssen, *Opt. Comm.* **62** (1987) 139; Z.Y. Ou, C.K. Hong, and L. Mandel, *Opt. Comm.* **63** (1987) 118; H. Fearn and R. Loudon, *Opt. Comm.* **64** (1987) 485.
16. N.G. Walker and J.E. Carroll, *Optical and Quantum Electronics* **18** (1986) 355; N.G. Walker, *J. Mod. Opt.* **34** (1987) 15.
17. H.J. Bernstein, *J. Math. Phys.* **15** (1974) 1677.
18. The method is to be found in M. Zukowski, *Phys. Lett.* **A177** (1993) 290.
19. S. Kochen and E.P. Specker, *J. Math. Mech.* **17** (1967) 59; J.S. Bell, *Rev. Mod. Phys.* **38** (1966) 447.
20. M. Redhead, *Incompleteness, Nonlocality and Realism* (Clarendon, Oxford; 1987).

TIME AS A DERIVED QUANTITY IN THE MICROMASER

MARLAN O. SCULLY
Texas A&M University, Department of Physics
College Station, Texas 77843

ABSTRACT

Aspects of the radiation-matter interaction in a cavity are reviewed. It is found that the concept of time appears as a natural result of phase shifts experienced due to the atom-field interaction.

Yakir Aharonov's contributions to, and love for, physics is an inspiration to us all. It is a pleasure, and an honor, to contribute this note to his Festschrift.

One of the cleanest and most interesting experiments in modern quantum optics involves resonant atoms passing through a high Q microwave cavity, i.e. the micromaser.[1,2] The "usual" treatment of the problem, in the notation of Fig. 1, assigns a time-of-flight $\tau = \ell/v$ to the atom-field interaction. In such a case, the Rabi oscillation between upper level and lower level, beginning with n photons in the cavity and an excited atom, i.e. beginning with $|\psi(0)\rangle = |a,n\rangle$, is described by

$$\psi(\tau) = \cos g\tau\sqrt{n+1}|a,n\rangle - i\sin g\tau\sqrt{n+1}|b,n+1\rangle \tag{1}$$

where we have assumed resonance between the atom and field, and g is the atom-field coupling constant.

Now there are several questions concerning Eq. (1), for example: What kind of center of mass wave function do we choose to yield the best approximation to Eq. (1)? Perhaps a "sharp" packet like $\delta(x-vt)$ so that the entrance and exit times are well defined, or perhaps a momentum state $\exp ipx$ so that the velocity is well defined, or perhaps some kind of Gaussian, minimum uncertainty, wave packet.

In order to address these, and other related points, we[3] were motivated to reconsider the simple problem of a plane wave center of mass wave function incident from the left as in Fig. 1

$$\Psi_{in}(t) = \exp ipx|a,n\rangle, \tag{2}$$

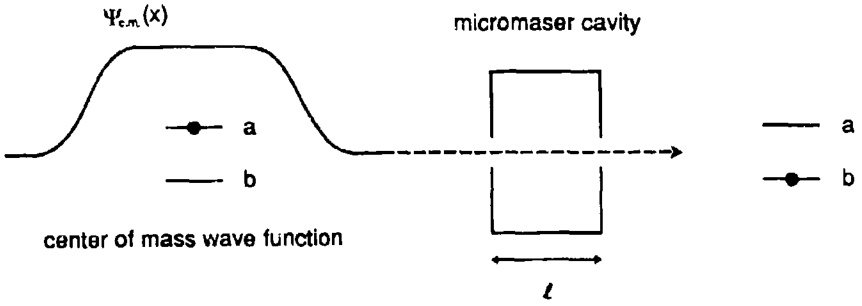


Fig. 1. Excited atom passes through cavity emerging in ground state.

interacting with a resonant cavity for which the Hamiltonian is

$$H = \frac{p^2}{2m} + g(x)[a\sigma^{\dagger} + \sigma a^{\dagger}], \tag{3}$$

where P is the c.m. momentum operator for atoms of mass m, $g(x)$ is the atom-field coupling constant which inside the cavity has a value g and vanishes outside, $a, a^{\dagger}$ are the annihilation and creation operators and $\sigma, \sigma^{\dagger}$ are the atomic lowering and raising operators.

Now the operator

$$\hat{\gamma} = \hat{a}\hat{\sigma}^{+} + \hat{\sigma}\hat{a}^{\dagger}, \tag{4}$$

has eigenstates

$$|\gamma_{\pm,n}\rangle = \frac{1}{\sqrt{2}}[|a,n\rangle \pm |b,n+1\rangle], \tag{5}$$

such that

$$\hat{\gamma}|\gamma_{\pm,n}\rangle = \pm\sqrt{n+1}|\gamma_{\pm,n}\rangle. \tag{6}$$

Therefore the Hamilton acting on states

$$\exp iP_{\pm} x|\gamma_{\pm,n}\rangle, \tag{7}$$

yields the following energies:

$$\frac{P^2}{2m} \qquad x < 0, \tag{8a}$$

$$\frac{P^2_{\pm,n}}{2m} \pm g\sqrt{n+1} \qquad 0 < x < \ell, \tag{8b}$$

$$\frac{P^2}{2m} \qquad x > \ell, \tag{8c}$$

and, therefore, by conservation of energy

$$\frac{P^2}{2m} = \frac{P^2_{\pm,n}}{2m} \pm g\sqrt{n+1}. \tag{9}$$

which implies

$$P_{\pm,n} = \left[P^2 \mp g\sqrt{n+1}/2m\right]^{1/2}, \tag{10}$$

and since the interaction energy is small compared to $P^2/2m$

$$P_{\pm} \cong P \mp g\sqrt{n+1}/(P/m). \tag{11}$$

Thus, for an excited state atom at $x = 0$, we have

$$\Psi(0) = |a,n\rangle = \frac{1}{\sqrt{2}}[|\gamma_{+,n}\rangle + |\gamma_{-,n}\rangle], \tag{12}$$

and at $x = \ell$ this becomes

$$\begin{aligned}\Psi(\ell) &= \frac{1}{\sqrt{2}}\left[e^{-iP_{+,n}\ell}|\gamma_{+,n}\rangle + e^{-iP_{-,n}\ell}|\gamma_{-,n}\rangle\right] \\ &= \frac{e^{-iP\ell}}{\sqrt{2}}\left[e^{ig\sqrt{n+1}\ell/(P/m)}|\gamma_{+,n}\rangle + e^{-ig\sqrt{n+1}\ell/(P/m)}|\gamma_{-,n}\rangle\right],\end{aligned} \tag{13}$$

hence, if we rewrite this back in terms of $|a,n\rangle$, $|b,n+1\rangle$ states we have

$$\begin{aligned}\Psi(\ell) = e^{-i\nu\ell}\Big[&\cos g\sqrt{n+1}\frac{\ell}{(P/m)}|a,n\rangle \\ &- i\sin g\sqrt{n+1}\frac{\ell}{(P/m)}|b,n+1\rangle\Big].\end{aligned} \tag{14}$$

This simple result has several interesting features, to wit:

1) The "correct" or "best" c.m. wave packet, in the sense of the question asked earlier, is a plane wave.

2) The atom should be thought of as being "spread" over the whole wave packet but is "somewhere" we just don't know where. That is, the c.m. wave packet may be many centimeters in extent but the atom is only a few angstroms in size. This result speaks to the proper interpretation of the wave packet in quantum mechanics. The relevance to the present problem is that we don't know when the atom enters the cavity, but when it does it acts like a point particle passing through in a "time" $\ell/(P/m)$.

3) Time here can be argued as being a derived quantity. We never introduce the concept of velocity. v_z and therefore the increment dt, but only the boost operator involving canonical momentum p. That is, we may think of our particle falling through a potential difference or being given an impulse so as to create the c.m. state e^{ipz}. Then the result (14) is obtained, which is to be compared with Eq. (1). The latter is, or course, derived from the *time* dependent Schrödinger equation.

4) The process of "photon emission" in the cavity as described by Eq. (14) involves well defined phase shifts (like $\exp(ig\sqrt{n+1}\ell/(P/m))$ which are just sufficient to ensure Rabi transitions. Note further that the vacuum Rabi angle $g\tau$ can be, and routinely is, $\pi/2$ in the experiments of Ref. 1.

5) This provides a natural basis for a which-path[3,4] detector.

6) The basis for enforcing complementarity, in the work of Refs. 3 and 4, is qualitatively different from the "randomization of phase" arguments made by Bohr in the classic Bohr-Einstein debate and by Furry and Ramsey in their discussion of the Aharonov Bohm effect.

Acknowledgements

The author wishes to thank the Office of Naval Research and the Walsh Foundation for their support and Yakir Aharonov for many stimulating and insightful discussions.

References

1. The first micromaser was built by the Max-Planck group of Prof. H. Walther, see D. Meschede, H. Walther and G. Müller, Phys. Rev. Lett. **54** (1985) 551.
2. For a review of cavity QED see S. Haroche and D. Kleppner, Physics Today **24** (1989) 24.

3. B.-G. Englert, J. Schwinger, A. O. Barut and M. O. Scully, Europhys. Lett. 14, 25 (1991).
4. M. O. Scully, B. G. Englert and H. Walther, Nature 351, 111 (1991).

TOWARD "IT FROM BIT"

The University of South Carolina

10-12 December 1992

John Archibald Wheeler
Physics Departments,
Princeton University and
University of Texas at Austin

I'm afraid that I'm one of the people whose primary goal is expressed by the Danish poet, Piet Hein, in his book of poems that he calls "Grooks":

"I'd like to know
What this show
Is all about
Before it's out."

And always in mind is that big question, "How come such a strange thing as existence?" And "How come the quantum?" And, in connection with the quantum, "How do we get the impression that there is *one* world out of the records of many observer-participators.?" If these questions verge on philosophy, then perhaps we can adopt as motto, "Philosophy is too important to be left to the philosophers." Among the philosophers, we have today two great schools--the Anglo-American and the Continental school. Heidegger, representative of the Continental school, in one of his books takes as central theme a passage from the German poet, Stefan George, "Without the word, no thing may be."

We have something a little like that in quantum theory. We know the photon does not exist in the atom before the act of emission and we know the photon does not exist in the detector after the act of detection. And we know that the passage of the photon from the atom to the detector is simply talk. So there is the great question: what is the role of the observer-participator in bringing about that which appears to be happening? Eugene Wigner at one time thought that consciousness was the key point, but I think that there were enough objections to that proposition from him and from others that he's given it up. We focus nowadays not on consciousness but on the act of detection or, better, what Niels Bohr describes as "an elementary quantum phenomenon. . .brought to a close by an irreversible act of amplification." The key point to my mind is expressed in the theme of "It from Bit." That is:

> Every "It", every particle, every field of force -- even the spacetime continuum itself -- derives its way of action, its very existence entirely, even if in some context indirectly, from the detector-elicited answers to yes or no questions, binary choices, bits.

In another way of wording the idea which I put up for examination, all things physical, all *its*, must in the end submit to an information-theoretic description.

The original Aharanov-Bohm experiment[1] (Fig. 1) illustrates this theme.

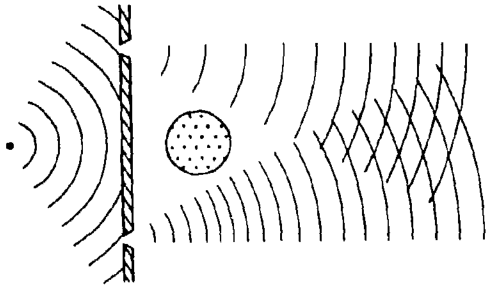


Figure 1. Idealized version of Aharonov-Bohm experiment. Point source of monoenergetic electrons. Electron waves emerge through the two slits in the diaphragm. They respond to the magnetic potential associated with the flux of magnetic lines of force that run throug;h the circular region embraced between but not touched by the two electron partial waves. The interference pattern at the right undergoes a shift which, expressed in count of fringes, measures directly the magnetic flux in appropriate Planck units.

Text continues:

Electrons emerge frcm the localized source at the left. Some p(netrate the double slit arrangement that divides the diagram. At the right, a flux of magnetic lines of force runs perpendicular to the ¬lane of the figure. They are e¬nbraced by the two branches of the electron beam but do not touch either one. Yet, as we kno./ from quantum mechanics, the equations recognize that the momentum is the sum of the kinetic momentum (proportional to electron wave number) and a potential momentum. The potential momentum is connected with the magnetic vector potential. The vector potential integrated around this circuit gives the amount of magnetic flux embraced by the circuit. What's more, the difference in wave number between below and above results in a shift of the interference fringes:

(Phase change around perimeter of the included area)
= 2π x (number of fringes shift of interference pattern)
= (electron charge) x (magnetic flux embraced) /hc.

We end up with the "bit" tally of fringe shift giving us directly the desired "it," the magnetic flux.

Another example of "it from bit" shows itself in quite another domain, the field of black-hole physics. Roger Penrose taught us about this marvelous process of interaction between an incoming object and a black-hole in which the two trade energy and angular momentum.[2] Demitrios Christodoulou, 19-year old graduate student who had never finished high school, got to work on analyzing these Penrose exchange processes. Yes, with their help, one can raise or lower the energy of a black hole and its angular momentum. But a certain combination of these two quantities, he found, can be raised, but never reduced.[3] (Figure. 2).

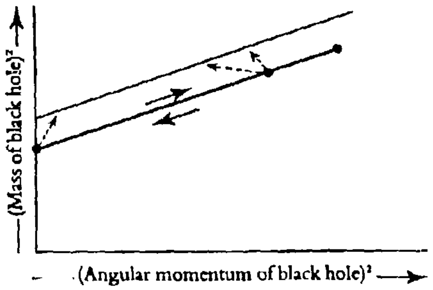


Figure 2. TRANSFORMATIONS:

Reversible ——> ; Irreversible --->

In a reversible transformation, the black hole stays on the Christodoulou line. An irreversible transformation takes the black hole off the line.

Text:

This combination is like entropy. Another graduate student, Jacob Bekenstein, came along and pointed out that this quantity not only is analogous to entropy, it must *be* entropy.[4]

I can recall confessing to Bekenstein how bad my conscience has always been in putting a hot teacup next to a cold one. Although energy is conserved in the exchange of heat between the two, that process increases the entropy of the world in an irreversible way that echoes unforgivingly down the corridors of time, forever. "But," I said to Jacob, "if a black-hole comes by, why can't I drop both teacups in and hide the evidence of my crime?" Bekenstein, however, is a man of great integrity. This proposed escape did not appeal to him. He came back a few months later with the conclusion that the black hole itself has entropy. How much entropy allows itself to be deduced out of the results of Christodoulou.

Well, we've all heard the lovely story about Brandon Carter bringing Bekenstein's paper to the attention of Stephen Hawking and the two of them deciding it was so preposterous they would write a paper to prove it was wrong. Then, in the course of the work, Stephen Hawking[5] finally found the formula for the emission of radiation by a black hole, and concluded the black hole *does have* temperature and the black hole *does have* entropy.

Then, going further in this domain, Kip Thorne and Wojciech Zurek analyzed a typical process in which particles and radiation fall into a black hole, and showed that the amount of information lost corresponds exactly to the increase in area measured in Bekenstein-Planck units.[6] So I sketched out a picture of the kind I wanted for a recent Scientific American Library book (Figure 3).

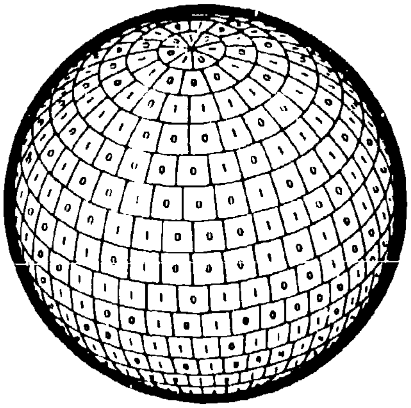


Figure 3. A totally symbolic representation of the Bekenstein-Hawking "Bit number" of a black hole. This number counts the number of boxes of Bekenstein size that can be pasted down (in imagination only!) on the horizon of the black hole. Each box contains one bit, one yes-or-no binary digit of information, about what went in to make that black hole.

Text:

It delighted me so tc hear how the two draftsmen created the picture. One of them drew the sphere and made these little boxes. The other threw coins, and, depending whether a head or a tail came up, put down a zero or a one.The enormous binary number one gets this way, with an enormous number of digits, does not **describe** the information but it **measures how many *bits*** of information.

It helps to think of the relevant information about what fell in as inscribed in a gigantic telephone book. Each page of the telephone book describes the energies of the photons and electrons and other entities that disappeared into the black hole in the act depicted. The pages in this telephone book we number in binary digits. The bit number of a black hole is only the number of the page in the telephone book, it's not a description of the information. The information needed is enormously more than this. So the black hole provides another example of the theme "it from bit."

Quantum theory and general relativity come together in many ways. The task that has long been on the books is so-called quantization of general relativity. But that phraseology of the task is misleading because it suggests that we ought to quantize spacetime. After all, isn't spacetime the overarching theme of general relativity? Actually, if there was anything that misled us all and prolonged the task more than anything else, it was reading and thinking that spacetime is the dynamic object. The dynamic object, however, is not spacetime. It is three-dimensional space geometry, and spacetime is the history of that geometry, or at least it's the classical history of space evolving in time. I like to consider a picture like Figure 4.

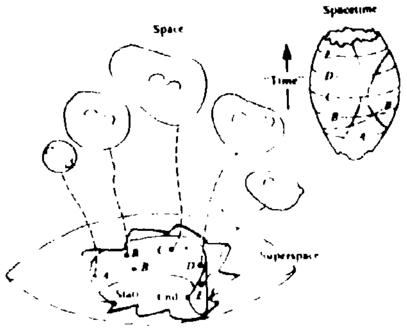


Figure 4.* Space, spacetime, and superspace. Upper left: Five sample configurations, A, B, C, D, E, attained by space in the course of its expansion and recontraction. Below: Superspace and those five sample configurations, each represented by a point in superspace. Upper right: Spacetime. A spacelike cut, like A through spacetime gives a momentary configuration of space. There is no compulsion to limit attention to a one-parameter family of slices, A, B, C, D, E through spacetime. The phrase "many-fingered time" is [illegible]

telling one not to so limit one's slices, and *B'* is an example of this freedom in action. The 3-geometries *B'* and *A, B, C, D, E,* like all 3-geometries obtained by all spacelike slices whatsoever through the given classical spacetime, lie on a single bent leaf of history, indicated in the diagram, and cutting its thin slice through superspace. A different spacetime, in other words, a different solution of Einstein's field equation, means a different leaf of history (not indicated) slicing through superspace. *From Charles W. Misner, Kip S. Thorne, and John A. Wheeler, Gravitation, W. H. Freeman & Co., New York, 1973, p. 1183.

Text:

At the upper right is spacetime, like an egg. A slice through that four-dimensional spacetime gives us 3-dimensional space. Thus A or B or B', etc. constitute a sequence of spacelike slices through spacetime. The two hazy curved lines symbolically depict two masses which interact and bend space in their vicinity but don't collide, and ultimately fly apart.

Quantum theory of spacetime leads us to think of a probability amplitude, not of a $\psi(x)$ as we do in the Schrödinger equation for a particle with one degree of freedom, but $\Psi({}^{(3)}G)$ as a functional of a 3-dimensional geometry: One probability amplitude for this 3-dimensional geometry, one for another and so on. Classical theory gives us a deterministic leaf of history, cutting through the space of all three geometries.

Quantum considerations teach us to speak of superspace. Superspace is an infinite dimensional manifold, each point which represents all the properties of a 3-dimensional geometry. So the 3-dimensional geometry, Δ, with all its lumps, bumps and wiggles, is symbolized by the point Δ in superspace and 3-geometry B is symbolized by another point. A different view of the same history, a different slice through spacetime, a slice that runs in another direction, let me call it B', is symbolized by another point in superspace. One is as good as another. No canonical choice.

But, there's one thing wrong with this classical picture: it gives us a deterministic leaf of history, sharply defined. Classically a certain 3-dimensional geometry is briefly realized or it isn't. Quantum mechanically, however, we know there is no such sharp yes or no distinction between 3-geometries. Instead, there's a 3-geometry-dependent classical probability amplitude that falls off sharply from this leaf of history that cuts through superspace. So the task of general relativity -- and it took a long time to recognize even what the task was in these terms -- was to find an equation for $\Psi\,({}^{(3)}\boldsymbol{G})$ and the equation can be written down[7] and made to look simple:

$$-\frac{\nabla^2\psi}{\left(\delta\,{}^{(3)}G\right)^2}+{}^{(3)}R\,\psi=0$$

in abbreviated form, or properly spelled out:[8]

$$\left(G_{ijkl}\frac{\delta}{\delta\gamma_{ij}}\frac{\delta}{\delta\gamma_{kl}}+\gamma^{1/2}\,{}^{(3)}R\right)\psi\left[{}^{(3)}G\right]=0$$

where:

$$G_{ijkl} \equiv \frac{1}{2}\gamma^{-1/2}(\gamma_{ik}\gamma_{jl}+\gamma_{il}\gamma_{jk}-\gamma_{ij}\gamma_{kl})$$

But to solve it directly has proved so far to be beyond our power.

Fortunately, Ashtekar[9], Smolin, Jacobson and Rovelli[10] and colleagues in a group of papers, more than 190 of them so far,[11] have given us a totally different way of dealing with the quantum theory which is closer to the "it from bit" point of view. One sees what it deals with most easily by going back to electromagnetism. There the simplest quantity to start with is the vector potential whose curl gives the magnetic field but whose integral around a circuit gives us, as in the Aharonov-Bohm experiment, the flux of magnetic field through the area embraced by that circuit. So one operation on the vector potential is differentiation and the other is integration around a closed circuit.[9,12]

Ashtekar, Smolin, Jacobson and Rovelli deal with a similar contrast between the usual differentiation of the connection in geometry that gives the curvature tensors on the one hand and integration around a circuit that gives a loop variable on the other hand and they come out with the conclusion that the probability amplitude in typical cases can be taken to depend only on the knot class of the loop.

It's enchanting to have a knot[13] come into the story because as we know, we can have knots in three dimensions, but not in four and not in two. So, it makes us a little more comfortable that we're right on the number of dimensions that we see around us.

So much for the formalism. But where does the concept of "it from bit" appear more directly in the work of Ashtekar, Rovelli and Smolin? It must be possible to express spacetime curvature by something that is analogous to the way we can get magnetic flux by a bit-like or fringe-count-measured integral around a circuit. And we can, so Jeeva Anandan teaches us.[12] Do an experiment where an uncharged particle -- preferably one with spin -- does for a loop in curved spacetime what the electron does in the Aharonov-Bohm experiment in a region where there is a magnetic field Count the shift of fringes and divide by the area encompassed between the two branches of the particle beam. In that way get the relevant component of the curvature tensor.[14] Repeat for three independent orientations. In this way we get the three components that define the relevant part of the Einstein curvature tensor.

Translate questions about physics into the counting of fringe shifts as a way to gain new insight, yes; that is the theme of "it from bit" in action. It leaves many old and unanswered questions still unanswered, but at least offers something closer to a formalism by which we someday might answer them: Does time necessarily end? Is the system necessarily closed in space, too? On the small scale, we know there must exist quantum fluctuations in the geometry. Are those fluctuations strong enough to give space everywhere, as I have argued,[7] a foam-like structure?

In a recent paper, Ashtekar, Rovelli and Smolin have made a further advance. They ask and answer a question. They ask, "What sorts of questions should we ask in order to get something that has a "bit"-like answer. They say if we have a loop that goes through space, and that loop cuts through a surface, then in the calculation, we get the contribution of a Planck area, an area that is the square, $\hbar G/c^3$, of the Planck length of 1899.

When I heard about Planck's 94 year-old paper, I was so fascinated by it, I looked it up. How come he could derive such a length so early before he or anyone had even the beginnings of quantum theory? The answer turned out to be motivation. He had set out early in life animated by the idea that so simple a feature of nature as black body radiation was surely a guide to something fundamental. He recognized then in the law of displacement of color of peak radiation with temperature something independent of all details of the structure of atoms and solids. Out of the constant in the displacement law he got a quantity which is essentially what subsequently came to be called later the Planck constant. Out of the Planck constant and the speed of light and the gravitational constant, he went on to show, one could form a complete set of units: space, time, mass, temperature and energy. And he urges that these quantities should serve for natural units in communications between people who live on different planets to break away from our own parochial Earth-based units. After all, we who live on this planet use a unit of time based on the turning of our own particular planet; a unit of mass based on the particular fluid we drink; and a unit of length based on the distance from Earth's equator to Earth's pole. In contrast, the Planck units are universal. I found it easy and attractive in 1954 to take Planck's paper and translate his general ideas into modern terminology and to give his units a name, the "Planck units," which he, of course, did not[7]. Ashtekar, Rovelli and

Smolin in an April 1993 preprint show that a loop drawn through space contributes to a certain integral one-Planck unit of area for any surface it intersects. Thanks to their work we have not only a working formalism but we have some new and interesting questions.

At the end of all mathematics, we come back to the question: How come there is any such thing as the quantum? And how much attention should we pay to someone like Heidegger who prior to the days of the quantum was so taken with the slogan of Stephan George, "Without the word, no thing may be."

I like to cite the game of Twenty Questions in its surprise version as suggestive. You come in the door and you start asking your questions. You know you only have twenty. Is it animal? "No." Is it vegetable? "No." Mineral? "Yes." Is it green? "No." Is it white? "Yes." You notice that the more questions you ask, the harder it is for your friends to answer them. They have to think and think and think. And finally, as the twenty questions are running out, you have to make up your mind to a definite word: Is it "cloud?" Your friend thinks and thinks and thinks and finally he says, "Yes." And everybody bursts out laughing and they explain that when you went out of the room they had agreed *not* to agree on a word. There was no word in the room when you came in. Everyone asked could answer your question as he wished, but with one small proviso. If you challenged and he couldn't provide a word compatible with his own answer and with all previous answers, he lost and you won. So it was just as difficult for everyone as it was for me. This game of Twenty Questions has a little of the flavor of the quantum theory of the electron in the atom. The electron, we sometimes think, has a position and a momentum in the atom, but no; not until we've made a measurement. We have to make up our mind what we're going to ask, but we can't ask both questions at once. Thus, the insertion of equipment to determine the one quantity automatically prevents us from installing and using such apparatus as would determine the other quantity.

So here we are at the end, and I'm still as puzzled as I was when I began with the questions, "How come existence? How come the quantum? " I don't know any more central question in all of physics than "How come the quantum?"

REFERENCES

1. Y. Aharonov and D. Bohm, "Significance of electromagnetic potentials in the quantum theory," *Phys. Rev. 115*, **3**, 3485-3491 (1950).

2. R. Penrose, "Gravitational collapse: The role of general relativity," *Riv. Nuovo Cimento* **1** (1969) 252.

3. D. Christodoulou, "Reversible and irreversible transformations in black-hole physics," *Phys. Rev. Lett.* **25** (1970) 1596-1597.

4. J. Bekenstein, "Black holes and the second law," *Nuovo Cimento Lett.* **4** (1972) 737-740 and "Generalized second law of thermodynamics in black-hole physics," *Phys. Rev. D.* **9** (1973) 3292-3300.

5. S. W. Hawking, "Black holes and thermodynamics," *Phys. Rev.* **13** (1976) 191-197.

6. W. Zurek and K. S. Thorne, "Statistical mechanical origin of the entropy of a rotating charged black hole," *Phy. Rev. Lett.***20** (1985) 2171-2175.

7. Wheeler, "On tho nature of quantum geometrodynamics." *Ann. of Phys.*2 (1957) pp. 604-614; J. A. Wheeler, "Superspace and the nature of quantum geometrodynamics," in Battelle Rencontres: 1967 Lectures in Mathematics and Physics, eds. C. M. DeWitt and J. A. Wheeler (New York, N.Y.: Benjamin, 1968) pp 242-307; reprinted as "Le superespace et la nature de la géométrodynamique quantique," in *Fluides et Champ Gravitationnel en Relativité Générale No. 170, Colloques Internationaux* (Éditions du Centre National de la Recherche Scientifique, Paris, 1969) pp. 257-322; C. W. Misner, K. S. Thorne and J .A. Wheeler, Gravitation (San Francisco, Calif: Freeman, 1973) §43.4 and p. 1217.

8. B. S. DeWitt, "Quantum theory of gravity," I:*Phys. Rev.* **160**, (1967)1113-1148; II: *Phys. Rev.***162**, (1967)1195-1239.

9. Abhay Ashtekar. Lectures on non-perturbative canonical gravity. Singapore: World Scientific (1991).

10. Carlo Rovelli. "Ashtekar's formulation of general relativity and loop-space non-perturbative quantum gravity: a report."*Class. Quan. Grav.*, **8** (1991) 1613-1675.

11. P. Hübner, updated by G. Gonzalez, Physics Dept., Syracuse Univ., June 1991, "Bibliography of publications related to classical and quantum gravity in terms of new canonical variables"

12. J. Anandan: "Quantum Interference and the Gravitational Field" in Quantum Theory and Gravitation, pp. 157–176, Edited by A. R. Marlow (Academic Press, 1980); J. Anandan: "Comment on geometric phase for classical field theories," *Phys. Rev. Lett.* **60**, 2555 (1988). See also Anandan, J. and Y. Aharonov, "Geometric quantum phase and angles," *Phys. Rev. D* **38**, 1863–1870 (1988), which includes references to the literature of the subject.

13. Louis H. Kauffman, Knots and Physics, Singapore: World Scientific (1991), and Michael Atiyah, The Geometry and Physics of Knots, Cambridge University Press (1990.

14. Arkady Kheyfets and Warner A. Miller, '"The boundary of a boundary principle in field theories and the issue of austerity of the laws of physics," in *J. Math. Phys.* **32** (11) (Nov. 1991) pp. 3168-3175, and A. Kheyfets and W. A. Miller, Oct. 1991. Carry a triad by Fermi-Walkertransport around a small loop of area $d\varepsilon$ and thereby get immadiately the Einstein tensor or, more vividly, the Élie Cartan moment of rotation; thus, in a nutshell:

$$g_{\alpha\beta}\ dT^{\alpha\beta} / \ d\varepsilon = G^{oo} \approx \textit{moment of rotation}$$

Here α designates the direction of the basis triad vector and β designates the orientation of the loop. See also Romano: Geometrodynamics versus Connection Dynamics, Ph.D. thesis, Syracuse University, 1990.

15. M. Planck, "Uber irreversible Strahlungsvorgänge," *Sitzungsber. Deut. Akad. Berlin, Kl. Math.-Phys. Tech.* (1899) 440-480.

SECTION 7

SPIN AND STATISTICS

(PARA)BOSONS, (PARA)FERMIONS, QUONS AND OTHER BEASTS IN THE MENAGERIE OF PARTICLE STATISTICS [1]

O.W. GREENBERG
Center for Theoretical Physics, Department of Physics, University of Maryland College Park, MD 20742-4111, USA,

D.M. GREENBERGER
Department of Physics, City College of the City University of New York New York, NY 10031, USA

and

T.V. GREENBERGEST
Department of Physics, Southern Methodist University, Dallas, TX 75275, USA

ABSTRACT

After some general comments about statistics and the TCP theorem, I discuss experimental searches for violations of the exclusion principle and theories which allow for such violations.

1. Introduction

It is a great pleasure to speak at this symposium honoring Yakir Aharonov. Because of the broad range of Yakir's interests, I have been able to see people who work in different areas than mine whom I don't usually see at conferences and to meet for the first time people whose names and work I know, but whom I had never had the opportunity to meet. Yakir is especially concerned with fundamental issues which have lasting interest, such as particle statistics. In the first part of my talk I will say some things about statistics and related issues which may not be generally known, and in the second part I will focus on how well we know that particles obey the statistics we think they obey and on theories which allow violations of statistics.

By way of introduction, I mention two relations involving spin which are on quite different footings. The relation between spin and isospin, that integer-spin particles have integer isospin and odd-half-integer-spin particles have odd-half-integer isospin, was suggested on the basis of few examples: the proton and neutron, which are in the odd-half-integer category and the three pions, which are in the integer category. Further, there was no fundamental basis for such a relation. When strange particles were discovered, this relation was found to be violated by the kaons, which have zero spin and isospin one-half, and by the lambda and sigma hyperons, which have spin one-half and integer isospin. Since there was no theory supporting this relation, it was easy to discard it. By contrast, the relation between spin and statistics first stated by Pauli[1] in 1936, that integer-spin particles obey Bose statistics and odd-half-integer-spin particles obey Fermi statistics was supported

[1] Talk presented by O.W. Greenberg

by many examples and, at least for free fields, was proved by Pauli from the basic requirement of local commutativity of observables. This relation has survived and is one of the most general results of quantum field theory.

2. General Comments about Statistics and Related Issues

2.1 Additivity of the Energy of Widely Separated Subsystems

The zeroth condition I discuss is the requirement that the energy of widely separated subsystems be additive. This requires that all terms in the Hamiltonian be "effective Bose operators" in that sense that

$$[\mathcal{H}(\mathbf{x}), \phi(\mathbf{y})]_- \rightarrow 0, |\mathbf{x}-\mathbf{y}| \rightarrow \infty. \qquad (1)$$

For example, $\mathcal{H}$ can't have a term such as $\phi(x)\psi(x)$, where ϕ is Bose and ψ is Fermi, because then the contributions to the energy of widely separated subsystems would alternate in sign. Such terms are also prohibited by rotational symmetry.

2.2 Statistics of Bound States is Determined by Statistics of Constituents

The well-known rule that a bound state of any number of Bosons and an even number of Fermions is a Boson, while a bound state with an odd number of Fermions is a Fermion, was first stated by Wigner,[2] who published in Hungarian and suffered the consequence of using a relatively inaccessible language. Later Ehrenfest and Oppenheimer[3] independently published this result in English.

2.3 Spin-Statistics Theorem

I distinguish between two theorems. The *physical* spin-statistics theorem is the theorem of Pauli mentioned above, local commutativity of observables requires that, given the choice between Bose and Fermi statistics, integer-spin particles must obey Bose statistics and odd-half-integer-spin particles must obey Fermi statistics. The phrase, *given the choice between*, is necessary, because the analogous connection holds between parabose or parafermi statistics and spin. The theorem which I prefer to call the spin-type-of-locality theorem, due to Burgoyne,[4] states that fields which commute at spacelike separation must have integer spin and fields that anticommute at spacelike separation must have odd-half-integer spin. Both the assumptions and the conclusions of the two theorems differ. The Pauli theorem explicitly assumes a choice between different types of *particle* statistics and concludes that if the wrong choice is made, then observables fail to commute at spacelike separation. For example, if one chooses Bose statistics for spin-one-half particles, i.e., uses Bose commutation relations for the annihilation and creation operators of the spin-one-half particles, then the commutator of the observables for the free theory will contain the $S^{(1)}(x-y)$ singular function, which does not vanish for spacelike $x-y$, rather than the $S(x-y)$ singular function which does. The theory (at least for the free case) still exists. The Burgoyne theorem makes no statement about particle statistics; rather it assumes a choice between *field* commutation rules. If the wrong choice is made, then the fields are identically zero, so the theory does not even exist. This latter theorem has a very general proof in the context of axiomatic field theory; however it says nothing about *particle* statistics.

2.4 Weakness of the TCP Theorem

In contrast to the spin-statistics theorem, which requires locality of observables, the TCP theorem holds regardless of locality, and is a much weaker theorem. Indeed, it is difficult to make a theory which violates TCP. This is clearly illustrated by Jost's example.[5] Jost shows that a free neutral scalar field whose annihilation and creation operators are quantized with anticommutation relations (and whose particles thus obey Fermi statistics) still obeys the *normal* TCP theorem. Cluster decomposition properties also hold regardless of the choice of commutation relations.

3. Search for Small Violations of Fermi and Bose Statistics

Now I come to the second part of my talk and discuss how to detect violations of Fermi or Bose statistics if they occur. Atomic spectroscopy is the first place to search for violations of the exclusion principle since that is where Pauli discovered it. One looks for funny lines which do not correspond to lines in the normal theory of atomic spectra. There are such lines, for example in the solar spectrum; however they probably can be accounted for in terms of highly ionized atoms in an environment of high pressure, high density and large magnetic fields. Laboratory spectra are well accounted for by theory and can bound the violation of the exclusion principle for electrons by something like 10^{-6} to 10^{-8}. A useful quantitative measure of the violation, V, is that V is the coefficient of the anomalous component of the two-particle density matrix; for fermions, the two-electron density matrix, ρ_2, is

$$\rho_2 = (1 - V)\rho_a + V\rho_s, \tag{2}$$

where $\rho_{a(s)}$ is the antisymmetric (symmetric) two-fermion density matrix. Thoma and Nolte,[6] in a contribution to a poster session here, discuss bounds on the violation of the exclusion principle for nucleons based on the absence of the nucleus ^5Li. Bounds also follow from the absence of ^5He. Mohapatra and I surveyed a variety of searches for violations of particle statistics in [7].

I will discuss an insightful experiment by Maurice and Trudy Goldhaber[8] which was designed to answer the question, "Are the electrons emitted in nuclear β-decay quantum mechanically identical to the electrons in atoms?" We know that the β-decay electrons have the same spin, charge and mass as electrons in atoms; however the Goldhabers realized that if the β-decay electrons were not quantum mechanically identical to those in atoms, then the β-decay electrons would not see the K shell of a heavy atom as filled and would fall into the K shell and emit an x-ray. The Goldhabers looked for such x-rays by letting β-decay electrons from a natural source fall on a block of lead. No such x-rays were found. The Goldhabers were able to confirm that electrons from the two sources are indeed quantum mechanically identical. At the same time, they found that any violation of the exclusion principle for electrons must be less than 5%.

Ramberg and Snow[9] developed this experiment into one which yields a high-precision bound on violations of the exclusion principle. Their idea was to replace the natural β source, which provides relatively few electrons, by an electric current, in which case Avogadro's number is on our side. The possible violation of the exclusion principle is that a given collection of electrons can, with different probabilities, be in different permutation symmetry states. The probability to be in the "normal" totally antisymmetric state would presumably be close to one, the next largest probability would occur for the state with its Young tableau having one row with two boxes, etc. The idea of the experiment is that each collection of electrons has a possibility of being in an "abnormal" permutation state. If the density matrix for a conduction electron together with the electrons in an atom has a projection onto such an "abnormal" state, then the conduction electron will not see the K shell of that atom as filled. Then a transition into the K shell with x-ray emission is allowed. Each conduction electron which comes sufficiently close to a given atom has an independent chance to make such an x-ray-emitting transition, and thus the probability of seeing such an x-ray is proportional to the number of conduction electrons which traverse the sample and the number of atoms which the electrons visit, as well as the probability that a collection of electrons can be in the anomalous state. Ramberg and Snow chose to run 30 amperes through a thin copper strip for about a month. They estimated the energy of the x-rays which would be emitted due to the transition to the K shell. No excess of x-rays above background was found in this energy region. Ramberg and Snow set the limit

$$V \leq 1.7 \times 10^{-26}. \tag{3}$$

This is high precision, indeed!

4. Theories of Violation of Statistics

4.1 Gentile's Intermediate Statistics

The first attempt to go beyond Bose and Fermi statistics seems to have been made by G. Gentile[10] who suggested an "intermediate statistics" in which at most n identical particles could occupy a given quantum state. In intermediate statistics, Fermi statistics is recovered for $n = 1$ and Bose statistics is recovered for $n \to \infty$; thus intermediate statistics interpolates between Fermi and Bose statistics. However, Gentile's statistics is not a proper quantum statistics, because the condition of having at most n particles in a given quantum state is not invariant under change of basis. For example, for intermediate statistics with $n = 2$, the state $|\psi\rangle = |k, k, k\rangle$ does not exist; however, the state $|\chi\rangle = \sum_{l_1, l_2, l_3} U_{k,l_1} U_{k,l_2} U_{k,l_3} |l_1, l_2, l_3\rangle$, obtained from $|\psi\rangle$ by the unitary change of single-particle basis, $|k\rangle' = \sum_l U_{k,l} |l\rangle$ does exist.

By contrast, parafermi statistics of order n is invariant under change of basis.[11] Parafermi statistics of order n not only allows at most n identical particles in the same state, but also allows at most n identical particles in a symmetric state. In the example just described, neither $|\psi\rangle$ nor $|\chi\rangle$ exist for parafermi statistics of order two.

4.2 Green's Parastatistics

H.S. Green[12] proposed the first proper quantum statistical generalization of Bose and Fermi statistics. Green noticed that the commutator of the number operator with the annihilation and creation operators is the same for both bosons and fermions

$$[n_k, a_l^\dagger]_- = \delta_{kl} a_l^\dagger . \tag{4}$$

The number operator can be written

$$n_k = (1/2)[a_k^\dagger, a_k]_\pm + \text{const}, \tag{5}$$

where the anticommutator (commutator) is for the Bose (Fermi) case. If these expressions are inserted in the number operator-creation operator commutation relation, the resulting relation is *trilinear* in the annihilation and creation operators. Polarizing the number operator to get the transition operator n_{kl} which annihilates a free particle in state k and creates one in state l leads to Green's trilinear commutation relation for his parabose and parafermi statistics,

$$[[a_k^\dagger, a_l]_\pm, a_m^\dagger]_- = 2\delta_{lm} a_k^\dagger \tag{6}$$

Since these rules are trilinear, the usual vacuum condition,

$$a_k|0\rangle = 0, \tag{7}$$

does not suffice to allow calculation of matrix elements of the a's and $a^\dagger$'s; a condition on one-particle states must be added,

$$a_k a_l^\dagger |0\rangle = \delta_{kl}|0\rangle . \tag{8}$$

Green found an infinite set of solutions of his commutation rules, one for each integer, by giving an ansatz which he expressed in terms of Bose and Fermi operators. Let

$$a_k^\dagger = \sum_{p=1}^{n} b_k^{(\alpha)\dagger}, \quad a_k = \sum_{p=1}^{n} b_k^{(\alpha)}, \tag{9}$$

and let the $b_k^{(\alpha)}$ and $b_k^{(\beta)\dagger}$ be Bose (Fermi) operators for $\alpha = \beta$ but anticommute (commute) for $\alpha \neq \beta$ for the "parabose" ("parafermi") cases. This ansatz clearly satisfies Green's relation. The integer p is the order of the parastatistics. The physical interpretation of p is that, for parabosons, p is the maximum number of particles that can occupy an antisymmetric state, while for parafermions, p is the maximum number of particles that can occupy a symmetric state (in particular, the maximum number which can occupy the same state). The case $p = 1$ corresponds to the usual Bose or Fermi statistics. Later, Messiah and I[11] proved that Green's ansatz gives all Fock-like solutions of Green's commutation rules. Local observables have a form analogous to the usual ones; for example, the local current for a spin-1/2 theory is $j_\mu = (1/2)[\bar{\psi}(x), \psi(x)]_-$. From Green's ansatz, it is clear that the squares of all norms of states are positive, since sums of Bose or Fermi operators give positive norms. Thus parastatistics gives a set of orthodox theories. Parastatistics is one of the possibilities found by Doplicher, Haag and Roberts[13] in a general study of particle statistics using algebraic field theory methods. A good review of this work is in Haag's recent book[14].

This is all well and good; however, the violations of statistics provided by parastatistics are gross. Parafermi statistics of order 2 has up to 2 particles in each quantum state. High-precision experiments are not necessary to rule this out for all particles we think are fermions.

4.3 The Ignatiev-Kuzmin Model and "Parons"

Interest in possible small violations of the exclusion principle was revived by a paper of Ignatiev and Kuzmin[15] in 1987. They constructed a model of one oscillator with three possible states: a vacuum state, a one-particle state and, with small probability, a two-particle state. They gave trilinear commutation relations for their oscillator. Mohapatra and I showed that the Ignatiev-Kuzmin oscillator could be represented by a modified form of the order-two Green ansatz. We suspected that a field theory generalization of this model having an infinite number of oscillators would not have local observables and set about trying to prove this. To our surprize, we found that we could construct local observables and gave trilinear relations which guarantee the locality of the current.[16] We also checked the positivity of the norms with states of three or less particles. At this stage, we were carried away with enthusiasm, named these particles "parons" since their algebra is a deformation of the parastatistics algebra, and thought we had found a local theory with small violation of the exclusion principle. We did not know that Govorkov[17] had shown in generality that any deformation of the Green commutation relations necessarily has states with negative squared norms in the Fock-like representation. For our model, the first such negative-probability state occurs for four particles in the representation of S_4 with three boxes in the first row and one in the second. We were able to understand Govorkov's result qualitatively as follows:[18] Since parastatistics of order p is related by a Klein transformation to a model with exact $SO(p)$ or $SU(p)$ internal symmetry, a deformation of parastatistics which interpolates between Fermi and parafermi statistics of order two would be equivalent to interpolating between the trivial group whose only element is the identity and a theory with $SO(p)$ or $SU(p)$ internal symmetry. This is impossible, since there is no such interpolating group.

4.4 Apparent Violations of Statistics Due to Compositeness

Before getting to "quons," the final type of statistics I will discuss, I want to interpolate some comments about apparent violations of statistics due to compositeness. Consider two ^3He nuclei, each of which is a fermion. If these two nuclei are brought in close proximity, the exclusion principle will force each of them into excited states, plausibly with small amplitudes for the excited states. Let the creation operator for the nucleus at location A be

$$b_A^\dagger = \sqrt{1-\lambda_A^2}\, b_0^\dagger + \lambda_A b_1^\dagger + \cdots, |\lambda_A| << 1 \tag{10}$$

and the creation operator for the nucleus at location B be

$$b_B^\dagger = \sqrt{1-\lambda_B^2}b_0^\dagger + \lambda_B b_1^\dagger + \cdots, |\lambda_B| << 1. \tag{11}$$

Since these nuclei are fermions, the creation operators obey fermi statistics,

$$[b_i^\dagger, b_j^\dagger]_+ = 0 \tag{12}$$

Then,

$$b_A^\dagger b_B^\dagger|0\rangle = [\sqrt{1-\lambda_A^2}\lambda_B - \lambda_A\sqrt{1-\lambda_B^2}]b_0^\dagger b_1^\dagger|0\rangle, \tag{13}$$

$$||b_A^\dagger b_B^\dagger|0\rangle||^2 \approx (\lambda_A - \lambda_B)^2 << 1, \tag{14}$$

so, with small probability, the two could even occupy the same location, because each could be excited into higher states with different amplitudes. This is not an intrinsic violation of the exclusion principle, but rather only an apparent violation due to compositeness.

4.5 "Quons"

Now I come to my last topic, "quons."[19] The quon algebra is

$$a_k a_l^\dagger - q a_l^\dagger a_k = \delta_{kl}. \tag{15}$$

For the Fock like representation which I consider, the vacuum condition

$$a_k|0\rangle = 0 \tag{16}$$

is imposed.

These two conditions determine all vacuum matrix element of polynomials in the creation and annihilation operators. In the case of free quons, all non-vanishing vacuum matrix elements must have the same number of annihilators and creators. For such a matrix element with all annihilators to the left and creators to the right, the matrix element is a sum of products of "contractions" of the form $\langle 0|aa^\dagger|0\rangle$ just as in the case of bosons and fermions. The only difference is that the terms are multiplied by integer powers of q. The power can be given as a graphical rule: Put o's for each annihilator and x's for each creator in the order in which they occur in the matrix element on the x-axis. Draw lines above the x-axis connecting the pairs which are contracted. The minimum number of times these lines cross is the power of q for that term in the matrix element.

The physical significance of q for small violations of Fermi statistics is that $q = 2\mathcal{V} - 1$, where the parameter $\mathcal{V}$ appears in Eq.( ). For small violations of Bose statistics, the two-particle density matrix is

$$\rho_2 = (1-\mathcal{V})\rho_s + \mathcal{V}\rho_a, \tag{17}$$

where $\rho_{s(a)}$ is the symmetric (antisymmetric) two-boson density matrix. Then $q = 1 - 2\mathcal{V}$.

For q in the open interval $(-1, 1)$ all representations of the symmetric group occur. As $q \rightarrow 1$, the symmetric representations are more heavily weighted and at

$q = 1$ only the totally symmetric representation remains; correspondingly, as $q \to -1$, the antisymmetric representations are more heavily weighted and at $q = -1$ only the totally antisymmetric representation remains. Thus for a general n-quon state, there are $n!$ linearly independent states for $-1 < q < 1$, but there is only one state for $q = \pm 1$. I emphasize something that many people find very strange: *there is no commutation relation between two creation or between two annihilation operators,* except for $q = \pm 1$, which, of course, correspond to Bose and Fermi statistics. Indeed, the fact that the general n-particle state with different quantum numbers for all the particles has $n!$ linearly independent states proves that there is no such commutation relation between any number of creation (or annihilation) operators. An even stronger statement holds: There is no two-sided ideal containing a term with only creation operators. Note that here quons differ from the "quantum plane" in which

$$xy = qyx \tag{18}$$

holds.

Quons are an operator realization of "infinite statistics" which were found as a possible statistics by Doplicher, Haag and Roberts[13] in their general classification of particle statistics. The simplest case, $q = 0$,[20], suggested to me by Hegstrom, was discussed earlier in the context of operator algebras by Cuntz.[21] It seems likely that the Fock-like representations of quons for $|q| < 1$ are homotopic to each other and, in particular, to the $q = 0$ case, which is particularly simple. Thus it is convenient, as I will now do, to illustrate qualitative properties of quons for this simple case. All bilinear observables can be constructed from the number operator, $n_k \equiv n_{kk}$, or the transition operator, n_{kl}, which obey

$$[n_k, a_l^\dagger]_- = \delta_{kl} a_l^\dagger, \quad [n_{kl}, a_m^\dagger]_- = \delta_{lm} a_k^\dagger \tag{19}$$

Although the formulas for n_k and n_{kl} in the general case[22] are complicated, the corresponding formulas for $q = 0$ are simple.[20] Once Eq.(18) holds, the Hamiltonian and other observables can be constructed in the usual way; for example,

$$H = \sum_k \epsilon_k n_k, \quad \text{etc.} \tag{20}$$

The obvious thing is to try

$$n_k = a_k^\dagger a_k. \tag{21}$$

Then

$$[n_k, a_l^\dagger]_- = a_k^\dagger a_k a_l^\dagger - a_l^\dagger a_k^\dagger a_k. \tag{22}$$

The first term in Eq.(22) is $\delta_{kl} a_k^\dagger$ as desired; however the second term is extra and must be canceled. This can be done by adding the term $\sum_t a_t^\dagger a_k^\dagger a_k a_t$ to the term in Eq.(21). This cancels the extra term, but adds a new extra term, which must be canceled by another term. This procedure yields an infinite series for the number operator and for the transition operator,

$$n_{kl} = a_k^\dagger a_l + \sum_t a_t^\dagger a_k^\dagger a_l a_t + \sum_{t_1, t_2} a_{t_2}^\dagger a_{t_1}^\dagger a_k^\dagger a_l a_{t_1} a_{t_2} + \ldots \tag{23}$$

As in the Bose case, this infinite series for the transition or number operator defines an unbounded operator whose domain includes states made by polynomials in the creation operators acting on the vacuum. (As far as I know, this is the first case in which the number operator, Hamiltonian, etc. for a free field are of infinite degree. Presumably this is due to the fact that quons are a deformation of an algebra and are related to quantum groups.) For nonrelativistic theories, the x-space form of the transition operator is[23]

$$\rho_1(\mathbf{x};\mathbf{y}) = \psi^\dagger(\mathbf{x})\psi(\mathbf{y}) + \int d^3z\psi^\dagger(\mathbf{z})\psi^\dagger(\mathbf{x})\psi(\mathbf{y})\psi(\mathbf{z})$$
$$+ \int d^3z_1d^3z_2\psi(\mathbf{z_2})\psi^\dagger(\mathbf{z_1})\psi^\dagger(\mathbf{x})\psi(\mathbf{y})\psi(\mathbf{z_1})\psi(\mathbf{z_2}) + \cdots, \tag{24}$$

which obeys the nonrelativistic locality requirement

$$[\rho_1(\mathbf{x};\mathbf{y}),\psi^\dagger(\mathbf{w})]_- = \delta(\mathbf{y}-\mathbf{w})\psi^\dagger(\mathbf{x}), \quad \text{and} \quad \rho(\mathbf{x};\mathbf{y})|0\rangle = 0. \tag{25}$$

The apparent nonlocality of this formula associated with the space integrals has no physical significance. To support this last statement, consider

$$[Qj_\mu(x), Qj_\nu(y)]_- = 0, \quad x \sim y, \tag{26}$$

where $Q = \int d^3xj^0(x)$. Equation (26) seems to have nonlocality because of the space integral in the Q factors; however, if

$$[j_\mu(x), j_\nu(y)]_- = 0, \quad x \sim y, \tag{27}$$

then Eq.(26) holds, despite the apparent nonlocality. What is relevant is the commutation relation, not the representation in terms of a space integral. (The apparent nonlocality of quantum electrodynamics in the Coulomb gauge is another such example.)

In a similar way,

$$[\rho_2(\mathbf{x},\mathbf{y};\mathbf{y}',\mathbf{x}'),\psi^\dagger(\mathbf{z})]_- = \delta(\mathbf{x}'-\mathbf{z})\psi^\dagger(\mathbf{x})\rho_1(\mathbf{y},\mathbf{y}') + \delta(\mathbf{y}'-\mathbf{z})\psi^\dagger(\mathbf{y})\rho_1(\mathbf{x},\mathbf{x}'). \tag{28}$$

Then the Hamiltonian of a nonrelativistic theory with two-body interactions has the form

$$H = (2m)^{-1}\int d^3x\nabla_x \cdot \nabla_{x'}\rho_1(\mathbf{x},\mathbf{x}')|_{\mathbf{x}=\mathbf{x}'} + \frac{1}{2}\int d^3xd^3yV(|\mathbf{x}-\mathbf{y}|)\rho_2(\mathbf{x},\mathbf{y};\mathbf{y},\mathbf{x}). \tag{29}$$

$$[H,\psi^\dagger(\mathbf{z}_1)\ldots\psi^\dagger(\mathbf{z}_n)]_- = [-(2m)^{-1}\sum_{j=1}^{n}\nabla^2_{\mathbf{z}_j} + \sum_{i<j}V(|\mathbf{z}_i-\mathbf{z}_j|)]\psi^\dagger(\mathbf{z}_1)\ldots\psi^\dagger(\mathbf{z}_n)$$
$$+\sum_{j=1}^{n}\int d^3xV(|\mathbf{x}\cdot\ \mathbf{z}_j|)\psi^\dagger(\mathbf{z}_1)\cdots\psi^\dagger(\mathbf{z}_n)\rho_1(\mathbf{x},\mathbf{x}'). \tag{30}$$

Since the last term on the right-hand-side of Eq.(30) vanishes when the equation is applied to the vacuum, this equation shows that the usual Schrödinger equation holds for the n-particle system. Thus the usual quantum mechanics is valid, with the sole exception that any permutation symmetry is allowed for the many-particle system. This construction justifies calculating the energy levels of (anomalous) atoms with electrons in states which violate the exclusion principle using the normal Hamiltonian, but allowing anomalous permutation symmetry for the electrons.[24]

I have not yet addressed the question of positivity of the squares of norms which caused grief in the paron model. Several authors have given proofs of positivity.[25-28] The proof of Zagier provides an explicit formula for the determinant of the $n! \times n!$ matrix of scalar products among the states of n particles in different quantum states. Since this determinant is one for $q = 0$. the norms will be positive unless the determinant has zeros on the real axis. Zagier's formula

$$det\ M_n(q) = \Pi_{k=1}^{n-1}(1 - q^{k(k+1)})^{(n-k)n!/k(k+1)}, \tag{31}$$

has zeros only on the unit circle, so the desired positivity follows. Although quons satisfy the requirements of nonrelativistic locality, the quon field does not obey the relativistic requirement, namely, spacelike commutativity of observables. Since quons interpolate smoothly between fermions, which must have odd half-integer spin, and bosons, which must have integer spin, the spin-statistics theorem, which can be proved, at least for free fields, from locality would be violated if locality were to hold for quon fields. It is amusing that, nonetheless, the free quon field obeys the TCP theorem and Wick's theorem holds for quon fields.[19]

It is well known that external fermionic sources must be multiplied by a Grassmann number in order to be a valid term in a Hamiltonian. This is necessary, because additivity of the energy of widely separated systems requires that all terms in the Hamiltonian must be effective Bose operators. I constructed the quon analog of Grassmann numbers[29] in order to allow external quon sources. Because this issue was overlooked, the bound on violations of Bose statistics for photons claimed in[30] is invalid.

4.6 Speicher's Ansatz

Speicher[27] has given an ansatz for the Fock-like representation of quons analogous to Green's ansatz for parastatistics. Speicher represents the quon annihilation operator as

$$a_k = \lim_{N\to\infty} N^{-1/2} \sum_{\alpha=1}^{N} b_k^{(\alpha)}, \tag{32}$$

where the $b_k^{(\alpha)}$ are Bose oscillators for each α, but with relative commutation relations given by

$$b_k^{(\alpha)} b_l^{(\beta)\dagger} = s^{(\alpha,\beta)} b_l^{(\beta)\dagger} b_k^{(\alpha)}, \alpha \neq \beta, \text{ where } \ s^{(\alpha,\beta)} = \pm 1. \tag{33}$$

This limit is taken as the limit, $N \to \infty$, in the vacuum expectation state of the Fock space representation of the $b_k^{(\alpha)}$. In this respect, Speicher's ansatz differs from Green's, which is an operator identity. Further, to get the Fock-like representation of the quon algebra, Speicher chooses a probabilistic condition for the signs $s^{(\alpha,\beta)}$,

$$\text{prob}(s^{(\alpha,\beta)} = 1) = (1 + q)/2, \tag{34}$$

$$\text{prob}(s^{(\alpha,\beta)} = -1) = (1-q)/2. \tag{35}$$

Rabi Mohapatra and I tried to get a specific ansatz for the $s^{(\alpha,\beta)}$ without success. I was concerned about that, but Jonathan Rosenberg, one of my mathematical colleagues at Maryland, pointed out that some things which are easy to prove on a probabilistic basis are difficult to prove otherwise. For example, it is easy to prove that with probability one any number is transcendental, but difficult to prove that π is transcendental. I close my discussion of Speicher's ansatz with two comments. First, one must assume the probability distribution is uncorrelated, that is

$$(1/N^2)\sum_{\alpha,\beta} s^{(\alpha,\beta)} = [(1/2)(1+q)(1) + (1/2)(1-q)(-1)] = q, \tag{36}$$

$$(1/N^3)\sum_{\alpha,\beta,\gamma} s^{(\alpha,\beta)} s^{(\beta),\gamma)} = q^2 \tag{37}$$

$$(1/N^3)\sum_{\alpha,\beta,\gamma} s^{(\alpha,\beta)} s^{(\beta,\gamma)} s^{(\gamma,\alpha)} = q^3, \tag{38}$$

etc. Secondly, one might think that, since Eq.(32) implies the analogous relation for two annihilators or two creators in the Fock-like representation, Speicher's ansatz would imply $a_k a_l - q a_l a_k = 0$, which we know cannot hold. This problem would arise if the ansatz were an operator identity, but does not arise for the limit in the Fock vacuum. Since a sum of Bose operators acting on a Fock vacuum always gives a positive-definite norm, the positivity property is obvious with Speicher's construction.

Speicher's ansatz leads to the conjecture that there is an infinite-valued hidden degree of freedom underlying q-deformations analogous to the hidden degree of freedom underlying parastatistics.

5. Acknowledgements

I thank the Graduate Research Board, University of Maryland, College Park, for a Semester Research Award and the University for a Sabbatical Leave during which this work was carried out. This work was supported in part by the National Science Foundation.

6. References

1. W. Pauli, *Ann. Inst. Henri Poincaré* **6** (1936) 137.
2. E.P. Wigner, *Math. und Naturwiss. Anzeiger der Ungar. Ak. der Wiss.* **46** (1929) 576.
3. P. Ehrenfest and J.R. Oppenheimer, *Phys. Rev.* **37** (1931) 333.
4. N. Burgoyne, *Nuovo Cimento* **8** (1958) 607.
5. R. Jost, *The General Theory of Quantized Fields*, (American Mathematical Society, Providence, 1965), pp 103-104.

6. M.H. Thoma and E. Nolte, *Phys. Lett. B* **291** (1992) 484.
7. O.W. Greenberg and R.N. Mohapatra, *Phys. Rev. D* **43** (1991) 4111.
8. M. Goldhaber and G.S. Goldhaber, *Phys. Rev.* **73** (1948) 1472.
9. E. Ramberg and G. Snow, *Phys. Lett. B* **2**[illegible] (1990) 438.
10. G. Gentile, *Nuovo Cimento* **17** (1940) 493.
11. O.W. Greenberg and A.M.L. Messiah, *Phys. Rev. B* **138** (1965) 1155.
12. H.S. Green, *Phys. Rev.* **90** (1953) 270.
13. S. Doplicher, R. Haag and J. Roberts, *Commun. Math. Phys.* **23** (1971) 199 and *ibid* **35** (1974) 49.
14. R. Haag, *Local Quantum Physics* (Springer, Berlin, 1992).
15. A.Yu. Ignatiev and V.A. Kuzmin, *Yad. Fiz.* **46** (1987) 786 [*Sov. J. Nucl. Phys.* **46** (1987) 444].
16. O.W. Greenberg and R.N. Mohapatra, *Phys. Rev. Lett.* **59** (1987) 2507.
17. A.B. Govorkov, *Teor. Mat. Fis.* **54** (1983) 361 [*Sov. J. Theor. Math. Phys.* **54** (1983) 234]; *Phys. Lett. A* **137** (1989) 7.
18. O.W. Greenberg and R.N. Mohapatra, *Phys. Rev. Lett.* **62** (1989) 712.
19. O.W. Greenberg, *Phys. Rev. D* **43** (1991) 4111.
20. O.W. Greenberg, *Phys. Rev. Lett.* **64** (1990) 705.
21. J. Cuntz, *Commun. Math. Phys.* **57** (1977) 173.
22. S. Stanciu, *Commun. Math. Phys.* **147** (1992) 211.
23. O.W. Greenberg, *Physica A* **180** (1992) 419.
24. G.W.F. Drake, *Phys. Rev. A* **39** (1989) 897.
25. D. Zagier, *Commun. Math. Phys.* **147** (1992) 199.
26. M. Bozėjko and R. Speicher, *Commun. Math. Phys.* **137** (1991) 519.
27. R. Speicher, *Heidelberg preprint no. 692* (1992). I thank Dr. Speicher for sending me his article prior to publication.
28. D.I. Fivel, *Phys. Rev. Lett.* **65** (1990) 3361; erratum, *ibid* **69** (1992) 2020. I have not checked that the erratum repairs the error in the proof of positivity in the article.
29. O.W. Greenberg, in *Workshop on Harmonic Oscillators, NASA Conference Pub. 3197*, ed. D. Han, Y.S. Kim and W.W. Zachary (NASA, Greenbelt, 1993).
30. D.I. Fivel, *Phys. Rev. A.* **43** (1991) 4913.

THE QUANT'UM THEORY OF MEASUREMENT AND THE FOUNDATIONS OF STATISTICAL MECHANICS

by

David Z Albert
Department of Philosophy
Columbia University
New York City

ABSTRACT

It is argued that certain recent advances in the construction of a theory of the collapses of Quantum-Mechanical wave functions suggest the possibility of an account of the tendencies of thermodynamic systems to approach their equilibrium states in which epistemic considerations play no role whatever.

0) Introduction

It is something of a cliche of theoretical physics, by now, to entertain the hope that the explicitly probablistic and explicitly time-reversal-asymmetric character of the collapses of quantum-mechanical wave-functions might somehow be related to, might somehow be *explanatory* of, the probablistic and time-reversal asymmetric character of the laws of thermodynamics.

And it is only slightly less of a cliche to point out that on second thought, on taking stock of precisely *what sorts* of probabilities and time-reversal-asymmetries collapses actually *exhibit*, the prospects for such an explanation don't look so good.

And what I want to do in this note is to rehearse the above considerations in some detail, and then to show how certain recent advances in our understanding of the collapse-process shed a radically different light on them.

1) What the Central Problem at the Foundations of Statistical Mechanics is

Recall some actual historical circumstance (of which you have witnessed a huge number) in which two macroscopic bodies whose temperatures initially differed were brought into thermal contact with one another, and in which those two bodies were not subsequently disturbed, and in which, ten minutes thereafter, the temperature-difference between those two bodies had decreased; and consider what the correct explanation of that decrease is.

The statistical-mechanical explanation of that decrease, both in the classical and in the quantum case, runs (crudely) like this:

The initial macrostate of that two-body system was compatible with a huge number of its possible microstates; and the overwhelming majority of those compatible microstates were ones which the deterministic equations of motions entail would evolve, over the next ten minutes, towards states in which the temperature-difference between those two bodies is smaller.

And

There's a principle of reasoning (which has gone under various names at various times: a principle of indifference, a principle of symmetry) to the effect that if we have no information bearing on the question of which one of a certain set of states obtains, then the probability we assign to any particular one of those states obtaining ought to be equal to the probability we assign to any particular other one of them obtaining.

And

As of the moment when those two bodies were brought together, nobody had any information whatever bearing on the question of which one of the above-mentioned compatible microstates of that system then obtained.

And so

As of the moment when those two bodies were brought together, everybody ought to have judged it to be overwhelmingly likely that the microstate of that system was one of those which the deterministic equations of motion entail would subsequently evolve towards states in which ıe temperature-difference between those two bodies is smaller.

And so

It was very much to be expected, as of the moment when those two bodies were brought together, that the temperatures of those two bodies would approach one another over the subsequent ten minutes.

And it is arguably the central problem at the traditional foundations of statistical mechanics that there has always seemed to be something manifestly unsatisfactory about that explanation.

It's something like this:

Nothing, surely, about what anybody may or may not have known about those two bodies at the moment when they were brought together can have played any role in bringing it about (that is:

in causing it to happen) that the temperatures of those two bodies subsequently approached one another! And so presumably nothing about what anybody may or may not have known about those two bodies at the moment when they were brought together can play any role in satisfactorily explaining why their temperatures subsequently approached one another. And yet (and this is what the trouble is\ the fact that nobody knew, as of the moment when they were brought together, precisely which one of the possible microstates of those two bodies then obtained plays a crucial role, an indispensable role, in the above so-called "explanation" of the fact that the their temperatures subsequently approached one another.

And what I want to do in this note is to describe how (notwithstanding the sorts of objections which were alluded to in the introduction, and which will be described more fully in the next section) certain recent developments in the quantum-mechanical theory of measurement suggest the possibility of a new and much improved foundation for statistical mechanics, in which no such trouble can arise.

2) What We're in Need of

Let's set up some notation.

Consider (again) the two-body system we talked about before. Call the set of those of the possible microstates of that system which are compatible with its initial macrostate "{C}". And call those microstates in {C} which the equations of motion entail will subsequently evolve towards states in which the temperature-difference between the two bodies is smaller "normal" microstates. And call those microstates in {C} which the equations of motion entail will subsequently evolve towards states in which the temperature-difference between the two bodies is bigger "abnormal" microstates.

And note that {C} will have a natural metric. In the classical case that metric will be the euclidian metric on the phase space, and in the quantum-mechanical case it will be the Hilbert-space metric generated by the absolute square of the inner product.

And note that a serviceable idea of what it amounts to for two microstates to be only microscopically different from one another, an idea of what it amounts to for two microstates to be within one another's microscopic neighborhoods, can be straightforwardly built out of that metric.

*

Now, it has already been mentioned here (and this is a fact that was made important use of in the above "explanation") that normal microstates in {C} vastly outnumber abnormal microstates in {C}; but it also happens to be the case (and this is a fact that was not made use of in the above "explanation") that normal microstates vastly outnumber abnormal microstates in every *individual microscopic neighborhood* of {C}[1], and (more particularly) that normal microstates vastly outnumber abnormal microstates even within the microscopic neighborhoods of every one of the *abnormal* the microstates in {C}.

And what that means is that the property of being a *normal* state is extraordinarily *stable* under small perturbations of those two bodies, and that the property of being an *abnormal* state is extraordinarily *unstable* under small perturbations of those two bodies.

And what *that* means is that if the two bodies we've been talking about here were in fact somehow being frequently and microscopically and randomly *perturbed*, *then* the temperatures of those two bodies would be overwhelmingly likely to approach one another *no matter which one* of the microstates in {C} initially obtained.

And so if the two bodies we've been talking about here *were*, in fact, somehow being frequently and microscopically and randomly perturbed, then the fact that their temperatures approached one another could be explained *objectively*, it could be explained (that is) *without reference to anything about what anybody happens to have known*.

And what I want to explore in this note is a way of taking advantage of that.

*

To begin with, a pair of perennial misunderstandings will need to be cleared up.

1) The perturbations in question here are going to have to be *genuinely random*, which is to say that they are going to have to be connected with *real physical chances* in the *fundamental laws of nature*.

That seems to have had a way of uncannily escaping people's attention. It has often been suggested in the literature, for example, that (since none of the macroscopic two-body systems of which we have ever had any experience, and none of the macroscopic two-body systems of which we ever *shall* have any experience, are genuinely *isolated* ones) those perturbations can be seen as arising simply from the interactions of the two-body system we've been talking about here with *its environment*.[2] But so long as whatever constitutes the environment of those two bodies is subject to the same sorts of deterministic laws as the constituents of those bodies themselves are, that sort of thing will patently get us nowhere: whatever perturbations arise from interactions with an environment like that will be "random" (if that's the word for it) only in the *explanatorily irrelevant* sense that *nobody happens to be aware of precisely what they are*.

2) Not just *any* real physical chances in the fundamental laws of nature will necessarily do the trick.

That's been missed too. That's what's been going on, for example, throughout the long tradition of attempts to connect the probabilities of statistical mechanics with the real physical chances in the fundamental laws of Quantum Mechanics.[3] What the trouble with all those attempts has always been (and this is the trouble that was alluded to in the introduction) is that on the standard way of thinking about Quantum Mechanics (that is: on the Copenhagen way of thinking about it, or on von Neumann's way of thinking about it) those chances only appear in connection with the act of measurement, and the tendency of a two-body system like the one we've been talking about to approach its equilibrium state presumably doesn't depend on anybody's having measured that system, or on anybody's being in the process of measuring that system, or on anybody's being about to measure that system, and so standard sorts of Quantum-Mechanical chances (even though they're real physical chances, and not merely epistemic ones) are presumably not the sorts of chances that can play any role whatever in explaining that tendency.

*

But it turns out that there are extremely good reasons (reasons which have been in the literature for an extremely long time, and which have nothing at all to do with the foundations of statistical mechanics) for believing that the standard way of thinking about quantum mechanics can't be right; and it turns out that a promising non-standard way of thinking about quantum mechanics exists in which chances come up somewhat differently.

That's what the next section will be about.

3) What We Have

Let me begin by very briefly rehearsing the quantum-mechanical measurement problem.

It comes up like this: Suppose that every physical system in the world invariably evolves in accordance with the linear deterministic quantum-mechanical equations of motion; and suppose that M is a good measuring instrument for a certain observable A of a certain physical system S. What it means for M to be a 'good' measuring instrument for A is just that for all eigenvalues a_i of A:

$$[\text{ready}>_M[A=a_i>_S \text{ ------> } [\text{indicates that } A=a_i>_M[A=a_i>_S \qquad (1)$$

where $[\text{ready}>_M$ is that state of the measuring instrument M in which M is prepared to carry out a measurement of A, '---->' denotes the evolution of the state of M+S during the measurement-interaction between those two systems, and $[\text{indicates that } A=a_i>_M$ is that state of the measuring instrument in which ,say, its pointer is pointing to the a_i- position on its dial. That is: what it means for M to be a 'good' measuring instrument for A is just that M invariably indicates the correct value for A in all those states of S in which A *has* any definite value.

The problem is that (1), together with the linearity of the equations of motion entails that:

$$[\text{ready}>_M \sum_i [A=a_i>_S \dashrightarrow \sum_i [\text{indicates that } A=a_i>_M [A=a_i>_S \qquad (2)$$

And that appears not to be what actually happens in the world. The right-hand-side of (2) is a *superposition* of various different outcomes of the A-measurement (and not any particular one of them), but what actually *happens* when we measure A on a system S in a state like the one on the left-hand-side of (2) is that *one* or *another* of those particular outcomes *does* emerge!

*

And so there's been a tradition of thinking that there *must*, in fact, be physical processes which do *not* proceed in accordance with the linear equations of motion: there has been a tradition of thinking that there must be such things in the world as non-linear, chance-governed, *collapses* of the wave-function.

And those collapses must somehow be connected with the act of *measurement*. But *how* connected, exactly?

The *standard* way of thinking about quantum mechanics connects them by *fiat*. It amounts to a *fundamental physical law*, on the standard way of thinking, that measurements *cause* collapses.[4]

But it's been understood for a long time that (since the meaning of a word like "measurement" is simply not *precise* enough to appear in any *fundamental physical law*, and since there isn't any plausible means of *making* it that precise) the standard way of thinking about that stuff can't possibly be the *right* way of thinking about it.

And so it's been understood for a long time tnat (if the argument just under equation 2 is accepted) there is going to have to be some sort of a bona fide physical theory of the collapse of the wave-function; of which the connection between collapses and measurements will be an approximate consequence, as opposed to a fundamental postulate.

*

Ghirardi, Rimini, and Weber have recently proposed a theory (the first theory) like that. Their idea (which is formulated for nonrelativistic quantum mechanics) goes like this: The wave function of an N particle system

$$\psi(r_1 \ldots . r_N, t) \tag{3}$$

usually evolves in accordance with the Schrodinger equation; but every now and then (once in something like $10^{15}/N$ seconds), at random, but with fixed probability per unit time, the wave function is suddenly multiplied by a normalized Gaussian (and the product of those two separately normalized functions is multiplied, at that same instant, by an overall renormalizing constant). The form of the multiplying Gaussian is:

$$K \exp[-(r-r_k)^2/2\Delta^2] \tag{4}$$

where r_k is chosen at random from the arguments r_n, and the width of the Gaussian, Δ, is of the order of 10^{-5} cm.. The probability of this Gaussian being centered at any particular point r is stipulated to be proportional to the absolute square of the inner product of (3) (evaluated at the instant just prior to this 'jump') with (4). Then, until the next such 'jump', everything proceeds as before, in accordance with the Schrodinger equation. The probability of such jumps per particle per second (which is taken to be something like 10^{-15}, as I mentioned above), and the width of the multiplying Gaussians (which is taken to be something like 10^{-5} cm.) are new constants of nature.

That's the whole theory. No attempt is made to explain the occurrence of these 'jumps'; that such jumps occur, and occur in precisely the way stipulated above, can be thought of as a new fundamental law; a beautifully straightforward and absolutely explicit *law of collapse*, wherein there is no talk at a fundamental level of 'measurements' or 'recordings' or 'macroscopicness' or anything like that.

Note that for isolated microscopic systems (i.e. systems consisting of small numbers of particles) 'jumps' will be so rare as to be completely unobservable in practice; and has been chosen large enough so that the violations of conservation of energy which those jumps will necessarily produce will be very very small (over reasonable time-intervals), even for macroscopic systems.

Moreover, if it's the case that every measuring instrument worthy of the name has got to include some kind of a *pointer*, which indicates the outcome of the measurement, and if that pointer has got to be a macroscopic physical object, and if that pointer has got to assume macroscopically different spatial positions in order to indicate different such outcomes (and all of this seems plausible enough, at least at first)[6], then the GRW theory can apparently guarantee that all measurements have *outcomes*.

Here's how: Suppose that the GRW theory is true. Then, for measuring instruments (M) such as were just described, superpositions like

[A>[M indicates that 'A'> + [B>[M indicates that B> (5)

(which will invariably be superpositions of macroscopically different localized states of some macroscopic physical object) are just the sorts of superpositions that don't last long. In a very short time, in only as long as it takes for the pointer's wave-function to get multiplied by one of the GRW Gaussians (which will be something of the order of $10^{15}/N$ seconds, where N is the number of elementary particles in the pointer) one of the terms in (5) will disappear, and only the other will propagate. Moreover, the probability that one term rather than another survives is (just as standard Quantum Mechanics dictates) proportional to the fraction of the norm which it carries.

*

The reader will already have guessed what all this has to do with the considerations of sections 1 and 2.

Let's make it explicit:

The suggestion is that *every single one* of the microstates in {C} (and not merely a large majority of them) will be overwhelmingly likely, on any theory of the collapse of the wave function like the one just described, to evolve, over the subsequent ten minutes, into states in which the temperature-difference between the two bodies is smaller.

The suggestion (that is) is that the 'jumps' in the theory just described are precisely the sorts of 'perturbations' we found ourselves in need of before, the ones whereby the time-irreversibility of the behaviors of macroscopic physical systems can be explained objectively, the ones whereby (as a matter of fact) it seems reasonable to hope that epistemic probabilities can be eliminated from physical science altogether.

The business of deciding whether or not to take this suggestion seriously will presumably involve detailed quantitative examinations of a host of particular cases; but there are reasons for being optimistic, even now, about how those examinations will come out. The point to bear in mind (and this is more or less what the point of this whole note is) is that the radical instability of the property of being an 'abnormal' microstate will entail that any one of an enormous selection of different perturbations will be capable of getting the job done. It would seem that we need only take care to insure that the perturbations in question be genuinely random (unlike in the 'environmental' scenarios), and that they be frequent and microscopic (unlike in the standard quantum-mechanical scenarios); and of course all of that gets taken care of for us in theories like the Ghirardi, Rimini, and Weber's.

*

Nonetheless, all of this will no doubt strike many readers as suspiciously neat. Maybe it will be useful, then, to finish up by briefly confronting what I suspect will turn out to be a typical sort of objection. This one was brought to my attention by Philip Pearle.

It goes like this: Consider an extraordinarily tiny gas, one which consists of something on the order of 10^5 molecules. Even gasses as tiny as that are known to be very likely to spread out (if space is available) over reasonable intervals of time, and yet gasses as tiny as that very unlikely to suffer even a single GRW-type collapse over such an interval, and so an explanation of the tendencies of gasses like that to evolve like that over intervals like that in terms of collapses of the wave-functions of their constituents is patently out of the question.

What the correct explanation of those tendencies will need to appeal to, I suspect, are collapses of the wave-functions of the microscopic constituents of the containers of those gasses.

And so the collapse-driven statistical mechanics that this note is about will entail that an extraordinarily tiny and extraordinarily compressed and absolutely isolated gas will have no lawlike tendency whatever to spread out.

And it can hardly be denied that that runs strongly counter to our intuitions.

What it does not run counter to, however (and this is what has presumably got to be important, in the long run), is our empirical experience.

References

1) "Microscopic" differences between states are to be understood here, by the way, as differences which are much smaller (and not merely smaller) than macroscopic ones; so that a set like {C} will necessarily contain a great number of non-overlapping microscopic neighborhoods.

2) See, for example, J. M. Blatt, Prog. Theor. Phys., 22, 745, 1959.

3) See, for example, the discussions in J. von Neumann, Mathematical Foundations of Quantum Mechanics (transl. by R. T. Beyer, Princeton University Press, Princeton, 1955) and D. Bohm, Quantum Theory (Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1951).

4) See, for example, P. A. M. Dirac, The Principles of Quantum Mechanics (Oxford University Press, Oxford, 1930).

5) G. C. Ghirardi, A. Rimini, and T. Weber, Phys. Rev. D34,

470.

6) Some second thoughts about this (which seem to me to be embarrassing but not fatal to the GRW theory) are recorded in D. Albert, "On the Collapse of the Wave Function" in Sixty-Two Years of Uncertainty, A. I. Miller (ed.) New York: Plenum Press, pp. 153-65.

TIME REVERSAL OF SPIN-SPIN COUPLINGS

A. PINES
University of California, Berkeley

ABSTRACT

It is still commonly believed that the decay of order in a coupled many-body system approaching thermodynamic equilibrium is irreversible. A famous example involves the decay of transverse magnetization from coupled nuclear spins in a solid. The free induction decay is analagous to the disappearance of order in a previously compressed gas diffusing to fill a larger container.

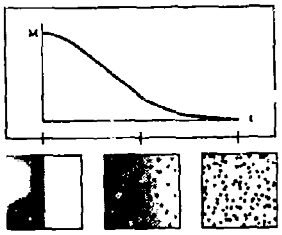


Figure 1: Decay of magnetization in a system of coupled spins, and analogy to diffusion in a lattice gas.

Of course it is recognized that, under unitary evolution, the order does not disappear, but evolves into subtle inter-particle correlations. The question is whether the initial order can be retrieved and, furthermore, whether the development of the correlations can be observed experimentally. I shall show examples of coupled many-body spin systems in which the apparently irreversibly decayed spin order is retrieved by time reversal of the spin-spin couplings under Haeberlen-Waugh averaging.

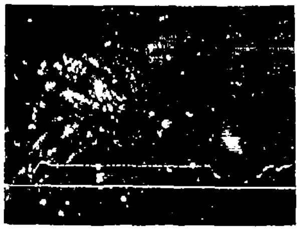


Figure 2: Experimentally observed decay of nuclear magnetization and multiple-pulse induced echo during the free induction of ^{19}F spins in solid calcium fluoride.

Unlike the Hahn spin echo in a system of uncoupled spins, which appears after a single refocussing pulse, time reversal in the coupled system requires a prolonged multiple-pulse sequence. Prior to time reversal the coupled system is, for most intents and purposes, in equilibrium and appropriately characterized by a canonical density operator. By means of coherent phase shifting, it is possible to detect and follow the time evolution of multiple-quantum spin coherences to high order (hundreds of particles).

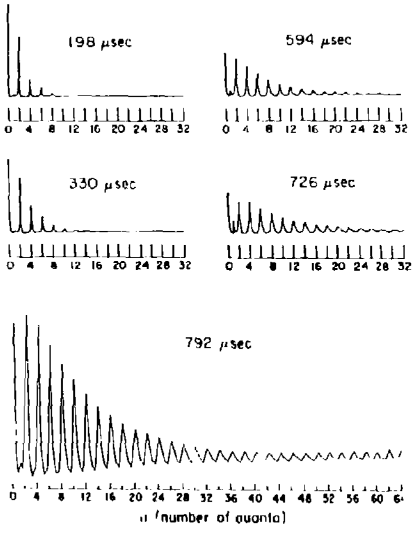


Figure 3: Experimentally observed evolution of multiple-quantum coherences arising from the development of correlations among coupled ^1H spins in solid hexamethylbenzene.

The multiple-quantum coherences reflect the existence of evolving spin-spin correlations (the subtle correlations that now sustain the original one-particle order). The coherences exhibit interesting examples of Abelian and non-Abelian geometric phases.

References: A. Pines, *NMR in Physics, Chemistry and Biology: Illustrations of Bloch's Legacy*, Proceedings of the Bloch Symposium, ed. W. A. Little, World Scientific (1990); J. W. Zwanziger, M. Koenig, and A. Pines, *Berry's Phase*, Ann. Rev. Phys. Chem., **41**, 601 (1990); L. Emsley and A. Pines, *Lectures on Pulsed NMR*, Proceedings of the CXXIII School of Physics "Enrico Fermi", ed. B. Maraviglia, North Holland (1994).

ELECTROMAGNETIC VELOCITY AND ACCELERATION OF NEUTRONS

by

C. R. Hagen
Department of Physics and Astronomy
University of Rochester
Rochester, NY 14627

and

J. Anandan
Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208

Abstract

The issue of neutron acceleration by uniform electromagnetic fields is examined. Although straightforward manipulation of the equations of motion implies such an effect, some doubt has continued to exist as to whether it is in principle observable. To resolve this matter a gedanken experiment is proposed and anlyzed using a wave packet construction for the neutron beam. By allowing arbitrary orientation for the neutron spin as well as for the electric and magnetic fields a nonvanishing acceleration of the center of the neutron wave packet is found which confirms the predictions of the canonical formalism. It is also shown that the difference between the canonical and kinetic momenta is in principle observable and in agreement with what one obtains using operator methods.

Although the motion of a neutral nonrelativistic spin-1/2 particle in a magnetic field is generally regarded as well understood, the corresponding problem with combined electric and magnetic fields continues to evoke discussion of underlying principles. Such a system can be described by the Hamiltonian

$$H = \frac{1}{2m} p^2 - \mu \cdot [\mathbf{B} - \frac{1}{mc} \mathbf{p} \times \mathbf{E}] \tag{1}$$

where $\mathbf{p}$ and m refer to the momentum and mass respectively of the particle (e.g., a neutron). The magnetic moment is $\mu = \gamma\sigma/2$ with σ being the set of Pauli spin matrices. It is assumed that the fields $\mathbf{E}$ and $\mathbf{B}$ are uniform and time independent. Upon calculating the commutators of H with $\mathbf{p}$, $\mathbf{x}$, and σ one finds

$$\dot{\mathbf{x}} = \mathbf{p}/m + \frac{1}{mc} \mathbf{E} \times \mu \tag{2}$$

$$\dot{\mathbf{p}} = 0 \tag{3}$$

$$\dot{\sigma} = \gamma\sigma \times [\mathbf{B} - \frac{1}{mc} \mathbf{p} \times \mathbf{E}] \quad . \tag{4}$$

It follows upon insertion of (4) into the time derivative of (2) that[1]

$$\ddot{\mathbf{x}} = \frac{\gamma}{mc} \mathbf{E} \times (\mu \times \mathbf{B}) + O(E^2) \quad . \tag{5}$$

Thus the inclusion of an electric field is seen to lead to a nonvanishing acceleration proportional in lowest order to both $|\mathbf{E}|$ and $|\mathbf{B}|$. It also implies that the canonical momentum $\mathbf{p}$ does not coincide with the kinetic momentum $m\mathbf{v}$.

Although the overall consistency of the canonical formalism of quantum mechanics would appear to offer no alternative to the result (5), there has been a recent suggestion[2] that in fact the predicted acceleration is essentially unobservable. It should be noted, however, that at least under the assumption of constant fields such a question must be capable of being resolved unambiguously by a direct calculation based on the Schrödinger equation. In particular a neutron which passes from a field free region to one described by the Hamiltonian (1) can be viewed as being subject to a constant but spin dependent potential. Corresponding to the particular spin eigenvalues the potential barrier for fixed **p** is then either attractive or repulsive.

To settle this issue one can imagine carrying out the following experiment. A wave packet describing a neutron is allowed to propagate in the field free region $z < 0$ and to enter the uniform field region $z > 0$ at normal incidence. The coordinate system is chosen so that the center of the wave packet passes through the origin at time $t = 0$. It is also assumed that the neutron is totally polarized in the direction of the unit vector **n**. At a distance z_D from the origin a detector is placed which measures the transverse displacement of the beam as a function of z_D to arbitrary accuracy. Since there can be no possibility of carrying out such measurements without a transverse localization of the beam, it is evident that the wave packet must be spread in at least one of the two transverse momentum variables. Furthermore, a spreading in the z coordinate is required in order to allow a time of flight to be inferred from the detector position z_D.

Thus the spatial part of the wave function for $z < 0$, $t < 0$ is given by

$$\begin{aligned}\psi(z, x \cdot n') = \int \frac{dp_z dp \cdot n'}{(2\pi)^2}\, e^{i(p\cdot n')(x\cdot n') + ip_z z} \\ \exp\left\{ -i\, \frac{(p\cdot n')^2 + p_z^2}{2m}\, t \right\} f(p_z, p\cdot n')\end{aligned} \tag{6}$$

where **n**′ is a unit vector in the x, y plane. The momentum space wave function $f(p_z, p\cdot n')$ is taken to be an even function in $p\cdot n'$ which is peaked around the point $p\cdot n' = 0$, $p_z = k$. It is normalized by the condition

$$\int \frac{dp_z dp\cdot n'}{(2\pi)^2}\, |f(p_z, p\cdot n')|^2 = 1 \quad .$$

While a Gaussian form for $f(p_z, p\cdot n')$ would allow an explicit calculation of the wave function to be performed, it is not in fact required for this problem. It may also be noted that a localization of the wave function in the second transverse direction is possible as well, but is basically irrelevant to the result.

When the wave packet passes through the origin the usual reflection and transmission effects are encountered although one clearly is interested only in the transmitted part for the purpose of this work. Matching the function and its derivative at $z = 0$ one finds that the spin part is unaltered while the transmitted beam has a form identical to that given by (6) provided that p_z in the exponential $e^{ip_z z}$ is replaced by the momentum component $\tilde{p}_z$ appropriate to propagation in the nonzero field region. The latter is obtained from the equation

$$\frac{p_z^2}{2m} = \frac{\tilde{p}_z^2}{2m} - \mu \cdot \left[\mathbf{B} - \frac{1}{mc}\,\mathbf{p} \times \mathbf{E}\right] \tag{7}$$

where (to lowest nonvanishing order in $\mathbf{E}$) it is the vector $\mathbf{p}$ which appears on the right hand side of (7) in combination with $\mathbf{E}$. Clearly, one could proceed at this point by considering separately the two eigenmodes of propagation and determining the appropriate spin part of the wave function by defining a spin basis with respect to the direction of the vector $\mathbf{B} - \frac{1}{mc}\mathbf{p} \times \mathbf{E}$. Fortunately, a simpler and more elegant approach is possible which involves calculating $\tilde{p}_z$ as a matrix in the spin space and using the projector

$$P_n = \frac{1}{2}\,(1 + \sigma \cdot \mathbf{n})$$

to include the initial polarization state of the beam.

To the required order one finds from (7) that

$$\tilde{p}_z = p_z + \frac{m\gamma}{2p_z}\,\sigma \cdot \left[\mathbf{B} - \frac{1}{mc}\,\mathbf{p} \times \mathbf{E}\right] \quad .$$

This then leads to the evaluation of $< x \cdot n' >$ by the expression

$$\begin{aligned} < x \cdot n' > = \int d(x \cdot n')dz(x \cdot n') \int \frac{dp_z dp'_z d(p \cdot n')d(p' \cdot n')}{(2\pi)^4} \\ e^{i[(p \cdot n') - (p' \cdot n')]x \cdot n' + i(p_z - p'_z)z} \\ \exp\left[-i\,\frac{(p \cdot n')^2 + p_z^2 - (p' \cdot n')^2 - p_z'^2}{2m}\,t\right] \\ f(p_z, p \cdot n')f^*(p'_z, p' \cdot n')T \end{aligned} \tag{8}$$

where

$$\begin{aligned} T = \text{Tr}\,\exp\left[iz\,\frac{m\gamma}{2p_z}\sigma \cdot (\mathbf{B} - \frac{1}{mc}\,\mathbf{p} \times \mathbf{E})\right]\,\frac{1}{2}\,(1 + \sigma \cdot \mathbf{n}) \\ \exp\left[-iz\,\frac{m\gamma}{2p'_z}\,\sigma \cdot (\mathbf{B} - \frac{1}{mc}\,\mathbf{p}' \times \mathbf{E})\right] \quad . \end{aligned} \tag{9}$$

The evaluation of the trace is simplified by the observation that because of the antisymmetry of the integrand in Eq. (8) under the simultaneous change of sign of $x \cdot n'$, $p \cdot n'$, and $p' \cdot n'$ only those terms in T which are odd in $p \cdot n'$ and $p' \cdot n'$ can contribute. Upon writing

$$\exp\left[iz\,\frac{m\gamma}{2p_z}\,\sigma\cdot(\mathbf{B}-\frac{1}{mc}\,\mathbf{p}\times\mathbf{E})\right] \\ = C + i|\mathbf{B}-\frac{1}{mc}\,\mathbf{p}\times\mathbf{E}|^{-1}\sigma\cdot(\mathbf{B}-\frac{1}{mc}\,\mathbf{p}\times\mathbf{E})S$$

where

$$(C,S) \equiv (\cos,\sin)\left[\frac{m\gamma z}{2p_z}\,|\mathbf{B}-\frac{1}{mc}\,\mathbf{p}\times\mathbf{E}|\right]$$

and using a prime to denote the same quantities when p is replaced by p', it is first noted that the CC' terms make no contribution to the trace. To the desired order one thus finds for (9) the result

$$T = \frac{i}{|\mathbf{B}|^2}\,SS'\mathbf{n}\times\mathbf{B}\cdot\mathbf{E}\times(\mathbf{p}-\mathbf{p}')\,\frac{1}{mc} \\ +\frac{i}{2}\,(CS'+C'S)\,\frac{1}{|\mathbf{B}|}\,\mathbf{n}\cdot\mathbf{E}\times(\mathbf{p}-\mathbf{p}')\,\frac{1}{mc} \quad .$$

Upon inserting this into (8) it is observed that the canonical commutation relations imply that the combination $(x\cdot n')(\mathbf{p}-\mathbf{p}')$ becomes $i\mathbf{n}'$ so that (8) reduces to

$$< x\cdot n' > \; = \int \frac{dp_z d(p\cdot n')}{(2\pi)^2}\,|f(p_z,p\cdot n')|^2 \\ \left[-\frac{1}{mc|\mathbf{B}|^2}\,S^2(\mathbf{n}\times\mathbf{B})\cdot(\mathbf{E}\times n')-\frac{1}{mc|\mathbf{B}|}\,CS\mathbf{n}\cdot\mathbf{E}\times\mathbf{n}'\right] \quad . \tag{10}$$

To complete the calculation one notes that it is sufficient to work to lowest order in $\mathbf{E}$ and $\mathbf{B}$ and to neglect corrections of the order of $\Delta p_z/k$ where Δp_z is the wave packet width in momentum space. This allows C and S to be replaced by 1 and $m\gamma z|\mathbf{B}|/2k$, respectively, thereby yielding for (10)

$$< x\cdot n' > = \{\frac{1}{2}\,t^2\,\frac{\gamma}{mc}\,\mathbf{n}'\cdot[\mathbf{E}\times(\gamma\,\frac{1}{2}\,\mathbf{n}\times\mathbf{B})] \\ + t\mathbf{n}'\cdot\mathbf{E}\times\frac{1}{2}\,\gamma\mathbf{n}\,\frac{1}{mc}\} \tag{11}$$

where z has been replaced by z_D which in turn is related to the time of flight t by $z_D = kt/m$. This identification is made possible by virture of the fact that the wave packet is localized in the z coordinate while its center moves with velocity k/m.

The verification of the expressions derived in ref. 1 for the acceleration and the kinetic momentum is now immediate. For short times t Eq. (5) clearly implies that there should exist a t^2 term in the mean transverse displacement whose coefficient is one-half the acceleration. It is striking that the calculation presented here yields an acceleration which has precisely the vector structure implied by the canonical formalism. Also noteworthy is the fact that the second term in (11) is linear in t and corresponds to the uniform drift of the particle beam implied by the difference between the canonical and kinetic momenta as indicated in Eq. (2). This term (unlike the acceleration) requires no magnetic field and one might therefore expect that it would offer fewer obstacles to an experimental detection.

One of us (CRH) acknowledges the support of U.S. Department of Energy Grant No. DE-FG02-91ER40685.

References

1. J. Anandan, Phys. Lett. A **138**, 347 (1989); J. Anandan in Proc. 3rd Int. Symp. Found. of Quant. Mech., Tokyo, 1989, edited by S. Kobayashi *et al.* (Physical Society of Japan, 1990).
2. R. C. Casella and S. A. Warner, Phys. Rev. Lett. **69**, 1625 (1992).

SECTION 8

QUANTUM REALITY AND PHENOMENOLOGY

A New Formulation of Quantum Mechanics

Yakir Aharonov

Department of Physics and Astronomy, University of South Carolina
Columbia, South Carolina 29208

and

School of Physics and Astronomy, Tel Aviv University
Tel Aviv, Israel 69978

Abstract

It is shown that the Schrödinger idea that considers a particle as an extended wave function is not wrong as is usually thought. The argument relies on a new method of measurement – the protective measurement – which measures the Schrödinger wave without disturbing it. However, to avoid paradoxes we have also to accept a new formulation of quantum mechanics which is based on two state vectors instead of one, the usual (history) state evolving toward the future and a second backward evolving (destiny) state.

I. The Standard Interpretation of Quantum Theory

When Schrödinger proposed his wave equation, there was much argument about the physical meaning of the wave function. While Schrödinger believed that the wave function for a single particle represents an extended object that was really moving in space, Born suggested that the wave function of a single particle has only a probabilistic meaning. That is, any experiment looking at a single particle will find that particle at only one location, but will never see it as an extended object. Only if we have an ensemble of particles, can we see the full implication of the wave function. For an ensemble the quantity $\psi^*(x)\psi(x)$ is proportional to the probability of finding the particle at the point x. We are able to infer the extended nature of a single particle only indirectly, for example, by analyzing a two slit experiment.

There are three general arguments usually presented as to why we can never see the wave function of a single particle. These arguments seem convincing, but we will later show why they are misleading.

1. In the laboratory we never see an extended object. If we make a measurement of an electron, we will always see it as a point on a

photographic plate, or a single track in a cloud chamber. It will always appear as a localized object, never as an extended object.

2. The second argument appeals to unitarity. Suppose we have two possible wavefunctions in the Schrödinger representation, ψ_1 and ψ_2. These are two different descriptions in space since, in general, the two functions are not orthogonal vectors in the Hilbert space. Suppose we now say that there is a measurement that can distinguish between the states ψ_1 and ψ_2 which are not orthogonal. That means there exists a measuring device with some state ϕ such that if the system is in state ψ_1 the state of the measuring device will go to ϕ_1, and if the system is in state ψ_2 the measuring device state will go to ϕ_2. To be able to distinguish between the two results, we must have ϕ_1 and ϕ_2 orthogonal. However, this violates unitarity since the initial states were not orthogonal.

   The usual argument is to have a large number of particles described by the same wavefunction ψ. That is, we start with a set $\psi_1(x_1)$, $\psi_1(x_2)$, ..., $\psi_1(x_N)$ and a set $\psi_2(x_1)$, $\psi_2(x_2)$, ...,$\psi_2(x_N)$. Using these two ensembles, we can distinguish between the two states, since the scalar product between any ψ_1 and ψ_2 is less than 1. The scalar product between the states of two sufficiently large ensembles of particles is essentially zero. Once again the statistical interpretation seems to be indicated.

3. The last argument is the most important since it forces us to adopt the two-vector formulation. Suppose at time t, there is a quantum particle whose wavefunction is non-zero in a large region. Let us assume there is an experiment which can determine that the particle is spread over this large region. We do this experiment and soon afterwards we do the usual experiment and find the particle localized at one position. If we were studying a charged particle, huge currents must flow to conserve charge. Otherwise there would be another frame of reference where the charge is not conserved. Thus the wavefunction cannot collapse infinitely fast. There is no way that an extended object can suddenly become a localized one.

We would like to be able to observe the full wave function. The wavefunction obeys the Schrödinger equation which tells us we have a vector in Hilbert space which evolves in a deterministic fashion. All the mystery in quantum mechanics occurs because we are told that we cannot observe the wavefunction. What we can observe is not what is described by the mathematics. The connection between what can be observed in the laboratory and what is described by the Schrödinger equation is only probabilistic. It would be beautiful if we could see the wavefunction directly.

II. Protective Experiments — An Example

The main argument for the reformulation suggested here is that there are experiments which protect the wavefunction so we can measure the wavefunction without destroying it. We call such experiments, protective experiments. Shelly Glashow suggested calling these protective experiments "in vivo" experiments. This is in analogy with biological experiments which preserve the life in a cell of small living objects. We shall consider below an example in which the protection is due to energy conservation.

Suppose we are given a particle described by a known Hamiltonian with discrete, non-degenerate eigenstates. We are told that the particle is in a definite eigenstate and we are asked to measure its wavefunction. A particular example of this would be an electron in the ground state of a hydrogen atom. In the standard interpretation, we measure the energy of this state and say that this is all that can be known. However, quantum mechanics contains much more information than this. It tells us that there is a wavefunction at each position in space. This is an infinite amount of additional information for a single particle. We will now discuss how we can extract this information without disturbing the wavefunction.

Measurement in an ideal quantum-mechanical experiment has been described by von Neumann. We let $\mathbf{H}_0$ be the Hamiltonian of the free system. This could be the Hamiltonian of an electron in the atom where, for simplicity, we take the proton mass as infinitely large. We let $\mathbf{A}$ represent the quantity we wish to measure, and let q be a variable of the measuring device. Then, the Hamiltonian of the system is

$$\mathbf{H} = \mathbf{H}_0 + g(t)\, q\, \mathbf{A}. \qquad (1)$$

where $g(t)$ is an interaction parameter. We choose

$$g(t) = \frac{g_0}{2T}\, e^{-|t|/T} \qquad (2)$$

Here T is the effective time of the measurement and g_0 is a constant representing the strength of the coupling between the system and the measuring device.

There are two interesting limits. The first is the impulsive limit where we take T → 0 and the other is the adiabatic limit where we take T → ∞. The usual experiment is to take the impulsive limit in which the experiment lasts an extremely short time. In this case we can ignore $\mathbf{H}_0$, and the momentum conjugate to q will be changed by one of the eigenvalues of **A**. We are only able to get probabilities for this change and hence cannot measure the wavefunction.

In the adiabatic limit the experiment lasts a long time while the coupling between the measuring device and the particle becomes very weak and approaches zero. Surprisingly, even though the coupling goes to zero, we can still get information about the particle and we get this information without changing the wavefunction. Indeed, in the adiabatic limit the ground state wave function is the ground state of the full Hamiltonian during the full time of the measurement. The only thing that can change is the phase.

We will first look at an eigenstate of q and then at a superposition of eigenstates. For an eigenstate, the adiabatic limit becomes a normal perturbation problem. The energy goes to the original energy plus a correction that goes to zero, that is

$$\mathbf{E} = \mathbf{E}_0 + g(t)\, q\, \langle \mathrm{A} \rangle \qquad (3)$$

where $\langle \mathrm{A} \rangle$ is the expectation value of **A** calculated with the original wavefunction. Now $\mathbf{E} - \mathbf{E}_0 \rightarrow 0$ but the total phase accumulated is

$$\int \mathbf{E}(t)\, \mathrm{d}t = \int \mathrm{E}_0\, \mathrm{d}t + g_0\ q\langle \mathrm{A} \rangle. \qquad (4)$$

Since quantum theory is a linear theory, what is true for q as an eigenstate is true for a superposition of eigenstates. If we start with the measuring device in a superposition of q's there will be a different phase associated with each value of q. If p is the momentum conjugate to q, the change in p will be $\delta p = g_0 \langle \mathrm{A} \rangle$. So we can measure not only the eigenvalue of an operator, but the average of an operator in a given state.

We can also clearly make N simultaneous measurements with N measuring devices, each measuring a different $\mathbf{A}_\mathrm{n}$. The Hamiltonian in this case is

$$\mathrm{H} = \mathrm{H}_0 + \mathrm{g(t)} \sum_{\mathrm{n}=\mathrm{o}}^{\mathrm{N}} q_\mathrm{n} \mathbf{A}_\mathrm{n}. \qquad (5)$$

If we choose the set of variables A_n to be the projection operators in different regions of space, the results for each A_n will be proportional to $\psi^*(x_n)\ \psi(x_n)$ at this point x_n and the entire set measures $\psi^*(x)\ \psi(x)$ in its full glory in all space.

III. Refutation of the Three Arguments of Section I

We must now show why the three seemingly very convincing arguments of Section I were misleading.

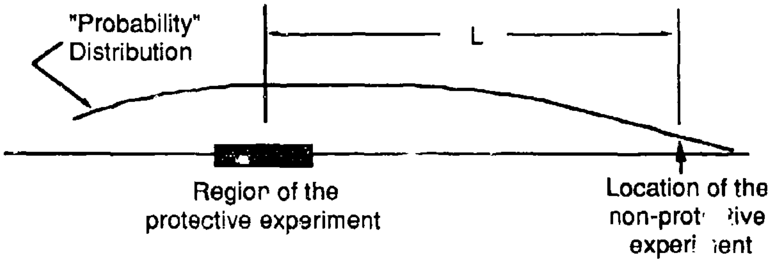
1. The first argument is easily discounted. The previous experiments were simply not the right experiments. Up to now, we have only designed experiments that would "kill" the wavefunction by looking for a localized particle. These were not "in vivo" experiments. In analogy with our biological example, if you do the wrong experiment on an organism, you will kill it. We previously were doing the wrong experiment on ψ and thus "killed" it.

2. The unitarity issue is resolved in an interesting way. The only states that were protected are nondegenerate eigenstates of the Hamiltonian. These eigenstates are orthogonal to each other, so no contradiction with unitarity arises. If we try to measure superpositions of two such states the system will collapse to one state or the other. We are still able to see the state in its full glory, but we see only one state out of the set of completely orthogonal nondegenerate eigenstates of the Hamiltonian. If we want to see other states such as a superposition of eigenstates, we must find a different protection since conservation of energy does not preserve them.

   The issue is not to think of measurement as just determining what we don't know. The real issue of measurement theory is determining what can manifest itself. If we have an electron passing through two slits, we can measure its wave function for a single particle and see the full glory of the interference spectrum. It is only necessary to devise the right protection.

3. Complete resolution of the third argument will be presented in the next section but it is interesting and fruitful to consider what happens, if while performing a protective measurement in one region of space, a usual position measurement is performed at some other location and finds the particle there. Can we violate causality and send signals faster than light? The answer is no. As an example suppose we have an

electron in the ground state of the hydrogen atom as shown in the figure.



We are doing our protective experiment in the vicinity of the proton and find the wave function density corresponding to the ground state. While we are doing our experiment some other physicist makes a non-protective measurement at a large distance L away from us and finds the whole particle there. This contradicts the outcome of our protective experiment. To avoid possibility of casual connection between these experiments, we must complete our protective experiment in a time T less than L/c. For finite time experiment we no longer can be sure that the electron remains in the ground state. There is a finite probability of exciting the state, which goes like e^{-ET}. This is the probability to make a mistake. On the other hand, the probability to find the particle at location L is

$$e^{-L\sqrt{2mE}}$$

where m is the mass of the particle and **E** is the binding energy. The only way to violate causality is to have a binding energy greater than $2mc^2$. We have a nice result. There is no way to consistently describe a single particle in relativistic quantum theory if the binding energy exceeds $2mc^2$.

IV. The Two-Vector Reformulation of Quantum Theory

To resolve violation of Lorentz covariance in the wavefunction collapse problem we must reformulate quantum mechanics. It is possible to do this by using the two-vector formulation. The two-vector formulation can be described as follows. Suppose we have a region of space where an experiment is performed. For example, in a scattering experiment we start with an incoming state, call it ψ_1 allow this prepared state to interact and produce a set of outgoing states corresponding to different outcomes. We want to select only particles that go into a particular outgoing state ψ_2. In classical physics, if we had a well-defined incoming state there would be only one outgoing state. In quantum mechanics there will be an ensemble of outgoing states. This allows us to define new quantum ensembles that do not have an analogy in classical physics. These are called pre-selected and post-selected ensembles. They are characterized by giving two boundary conditions on the particles. These are the boundary conditions at the start of the experiment and the boundary conditions at the end of the experiment. That is, if an incoming state ψ_1 can produce the following set of outgoing states ψ_2, ψ_3, etc., then we form separate ensembles for those experiments that produce the pair (ψ_1,ψ_2) and those that produce (ψ_1,ψ_3), etc.

This suggests characterizing each quantum particle in the pre-selected and post-selected ensemble by two states. Each quantum particle is described at any instant by two vectors that we will call the history vector and the destiny vector. This concept will enable us to explain how a distribution that was extended in space can suddenly be replaced by a distribution that is peaked near a given position.

What we measure is not $\rho(x) = \psi^*(x)\ \psi(x)$ but the density

$$\rho_{12}(x) = \frac{\psi^*_2(x)\ \psi_1(x)}{\int_{-\infty}^{\infty} \psi^*_2(x')\ \psi_1(x')\ dx'}$$

where ψ_1 is the history vector and ψ_2 is the destiny vector. In all protective experiments what is measured is not the average of either of these states but the above combination. In the usual non-protective experiments, the history vector and the destiny vector were the same, so this distinction was not obvious.

Let us consider again the paradoxical situation of the argument 3. Let the initial state ψ_1 be a superposition of the two localized states. The final state ψ_2 is one of these localized states. We might obtain it by just looking and not finding the particle in another place. The paradox is how the particle "jumps" instantaneously to the first location just by not observing it in the second location. The way out is that the particle was in the first location during the whole period between the two measurements! Indeed, the two vector density $\rho_{12}(x)$ is non-zero only when both ψ_1 and ψ_2 are not zero, i.e., only in the first location. Similarly we can resolve the problem of how an extended particle becomes localized. The product of an extended particle and a localized particle is always a localized particle. It is localized all the time.

We resolve argument 3 by thinking of a quantum system as being described by two vectors, the history vector and the destiny vector, rather than by one vector. We no longer violate causality since the description depends also upon what happens later, not just upon what has happened. If you change your mind about what you will measure, the destiny vector must be changed all the way back to the beginning just as we would have to change the history vector if we had decided to perform a different experiment. This is analogous to the Einstein-Podolsky-Rosen (EPR) experiment. In EPR, we have already learned that if we take a single system that is already correlated to another system, and make a measurement on one of the systems, it immediately changes the state of the other system. In an ensemble, the probability distribution remains unchanged, so we cannot use this to send information faster than light. In the same way here, the future state changes the present for an individual quantum system; but it doesn't change the probability distribution for an ensemble. Therefore, it cannot be used to transfer information backwards in time.

Conclusion

We have described a new type of experiment, the protective measurement, through which we can observe the extended wavefunction of a single particle in its full glory. This reality of the wavefunction strongly supports a new interpretation of quantum mechanics, the 2-vector formulation, in which there are 2 vectors describing a quantum system, the usual vector propagating from the past and a second one propagating backwards from the future. We show how this interpretation resolves the arguments given against the observability of the wavefunction of a single particle.

Acknowledgments

The following took part in developing this new formulation of quantum theory: Susan D'Amato, David Albert, Lev Vaidman, Sandu Popescu, Jeeva Anandan, Benny Resnik, and Daniel Rohrlich. This research was supported in part by grant U25/91-1 of the Basic Research Foundation (administrated by the Israel Academy of Science) and by NSF grant PHY-8807812.

MEASUREMENT OF NONLOCAL VARIABLES WITHOUT BREAKING CAUSALITY

LEV VAIDMAN
School of Physics and Astronomy
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University, Tel-Aviv, 69978 ISRAEL

ABSTRACT

We report results of an investigation of relativistic causality constraints on the measurability of nonlocal variables. We show that measurability of certain nondegenerate variables with entangled eigenstates contradicts the principle of causality, but that there are other, certainly nonlocal, variables which can be measured without breaking causality. We show that any causal measurement of nonlocal variables must erase certain local information. For example, for a system of two spin-1/2 particles, even if we take the weakest possible definition of verification measurement, verification of an entangled state must erase all local information.

1. Measuring Momentum of a Particle

As early as 1931, Landau and Peierls[1] showed that relativistic causality imposes new restrictions on the process of quantum measurement. Although some of their arguments were not precise, it was commonly accepted that we cannot measure instantaneously nonlocal properties without breaking relativistic causality.

The first example is the measurement of momentum of a particle. Consider a particle localized in a small region. Measurement of its momentum, irrespective of the outcome, will spread the particle all over the space. There will be a nonzero probability to find the particle at a very large distance from its original place immediately after the (instantaneous) momentum measurement, so it seems that the particle moves faster than light. However, this argument is not decisive. Relativistic causality states that it is impossible to send a *signal* with superluminal velocity. It does not forbid instantaneous measurement of momentum, say at $t = 0$. The instantaneous measurement interaction will take place all over the space and it can create particles everywhere. Thus, the probability of finding the particle at a given location after the momentum measurement might be independent of what we did to the particle located far away before the measurement. Therefore, the possibility of instantaneous momentum measurement does not lead automatically to the possibility of sending signals with superluminal velocity.

Nevertheless, if we can measure the momentum of a spin-1/2 particle without affecting its spin, then we can violate causality. Indeed, let us assume that we know that at time $t = 0$ the momentum measurement will be performed. At the time $t = -\epsilon$ we decide to prepare the state of the particle "up" or "down" according to the signal we want to send. Then we can measure the spin component of the particle which is detected at time $t = +\epsilon$ far from its original location and thus send information with superluminal velocity. (The probability of finding the particle at a given place is very small, but we can use a large ensemble of identical particles and thus we can build a reliable superluminal transmitter.)

2. Constraints on Nonlocal Measurements of Two Spin-1/2 Particles

Although momentum measurement is a basic problem, it is still not the simplest example we may consider. Significant progress in understanding causality constraints on quantum measurement was made by considering an even simpler example: measurements of spin variables of two spin-1/2 particles separated in space. This is the system on which Bohm and Aharonov[2] and later Bell[3] analyzed the EPR argument and reached far-reaching conclusions regarding the nonlocal structure of quantum theory.

In order to show how measurability of nonlocal variables contradicts relativistic causality let us consider an operator with the following nondegenerate eigenstates:

$$\begin{aligned}
|\psi_1\rangle &= |\uparrow\rangle_1 |\uparrow\rangle_2 \\
|\psi_2\rangle &= |\downarrow\rangle_1 |\downarrow\rangle_2 \\
|\psi_3\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 |\downarrow\rangle_2 + |\downarrow\rangle_1 |\uparrow\rangle_2) \\
|\psi_4\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 |\downarrow\rangle_2 - |\downarrow\rangle_1 |\uparrow\rangle_2)
\end{aligned} \tag{1}$$

This operator corresponds to a nonlocal variable because its eigenstates are nonlocal. We call the state of the composite system nonlocal when it cannot be represented as a product of states corresponding to localized parts of the system; these states are also known as *entangled* states.

Let us show that the measurability of this variable contradicts relativistic causality. To this end we perform the following set of measurements:

i) We prepare state $|\uparrow\rangle_2$ of particle number 2 a long time before the time $t = 0$.

ii) At time $t = -\epsilon$ we prepare state $|\uparrow\rangle_1$ or $|\downarrow\rangle_1$ of particle number 1 according to the message we want to send from particle 1 to particle 2.

iii) At time $t = 0$ we measure the variable defined by the nondegenerate eigenstates of Eq. (1).

iv) At the time $t = \epsilon$ we measure the spin component of particle 2.

The two events, choosing the spin of particle 1 and measurement of the spin of particle 2, are space like separated, and therefore must be causally disconnected. But if we choose spin "up" for particle 1, then the state of the composite system before the time $t = 0$ is $|\uparrow\rangle_1 |\uparrow\rangle_2$, the measurement at the time $t = 0$ does not change it (since it is an eigenstate), and thus the spin measurement of particle two will yield "up" with probability one. If, instead, at the time $t = -\epsilon$, we put, the particle 1 in the state "down," then the state of the composite system before the measurement (iii) is $|\downarrow\rangle_1|\uparrow\rangle_2$. This state is not one of the eigenstates of the nonlocal operator, and therefore the measurement at time $t = 0$ will change it. Since the scalar product between $|\downarrow\rangle_1|\uparrow\rangle_2$ and the eigenstates is not vanishing only for the eigenstates $|\psi_3\rangle$ and $|\psi_4\rangle$, the state after $t = 0$ will be one of those. But for both $|\psi_3\rangle$ and $|\psi_4\rangle$ the probability to find the spin "up" for particle 2 is just 1/2. We have shown that the possibility of measuring nonlocal variable described by eigenstates (1) allows us to change the probability of the result of a spin measurement performed on particle 2 by acting on particle 1 a time only 2ϵ before the measurement on particle 2; and since the distance between the particles might be larger than $2\epsilon c$, this procedure represents a superluminal signal transmitter.

3. Measurable Nonlocal Variables

The examples above may lead us to believe that measurement of any nonlocal variable breaks relativistic causality. This, in fact, was generally believed until Aharonov and Albert[4] found a method involving solely local interactions (hence consistent with the causality principle) which does allow us to measure certain nonlocal variables. In particular, we can measure the variable $\sigma_{1z} + \sigma_{2z}$. The method applies the standard von Neumann measuring procedure to a measuring device consisting of two parts which were prepared in an *entangled* state before the measurement. Each part of the measuring device interacts with one of the particles for a short time, and is observed immediately after by a local observer. The combined observations of the two observers (one at each particle) determines whether the state is $|\psi_1\rangle$, $|\psi_2\rangle$ or belongs to the subspace spanned by $|\psi_3\rangle$ and $|\psi_4\rangle$. The feature of this method is that while it measures $\sigma_{1z} + \sigma_{2z} = 0$, it does *not* measure the spin of each particle separately. The details of the method of nonlocal measurements can be found in Ref. (5).

It might seem that the measurability of the operator $\sigma_{1z} + \sigma_{2z}$ has something to do with its having a complete set of eigenstates which are not entangled. But this is not the explanation. The next example shows an operator with nondegenerate eigenstates that are all entangled but which is, nevertheless, measurable by local interactions. The eigenstates of the nondegenerate operator are

$$\begin{aligned}
|\psi_1\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\uparrow\rangle_2 + |\downarrow\rangle_1|\downarrow\rangle_2)\\
|\psi_4\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\uparrow\rangle_2 - |\downarrow\rangle_1|\downarrow\rangle_2)\\
|\psi_3\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\downarrow\rangle_2 + |\downarrow\rangle_1|\uparrow\rangle_2)\\
|\psi_4\rangle &= \frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\downarrow\rangle_2 - |\downarrow\rangle_1|\uparrow\rangle_2)
\end{aligned} \tag{2}$$

This operator can be measured[6] using a set of nonlocal operators with degenerate eigenstates (such as $\sigma_{1z} + \sigma_{2z}$), where the particles 1 and 2 are far from one another. Recently[7] the measurability of operators for two spin-1/2 particles has been analyzed, and it was shown that the only measurable nondegenerate operators are those with eigenstates of two possible types:

$$\begin{aligned}|\psi_1\rangle &= |\uparrow_z\rangle_1 |\uparrow_{z'}\rangle_2\\ |\psi_2\rangle &= |\uparrow_z\rangle_1 |\downarrow_{z'}\rangle_2\\ |\psi_3\rangle &= |\downarrow_z\rangle_1 |\uparrow_{z'}\rangle_2\\ |\psi_4\rangle &= |\downarrow_z\rangle_1 |\downarrow_{z'}\rangle_2\end{aligned} \tag{3a}$$

or

$$\begin{aligned}|\psi_1\rangle &= \frac{1}{\sqrt{2}}(|\uparrow_z\rangle_1 |\uparrow_{z'}\rangle_2 + |\downarrow_z\rangle_1 |\downarrow_{z'}\rangle_2)\\ |\psi_2\rangle &= \frac{1}{\sqrt{2}}(|\uparrow_z\rangle_1 |\uparrow_{z'}\rangle_2 - |\downarrow_z\rangle_1 |\downarrow_{z'}\rangle_2)\\ |\psi_3\rangle &= \frac{1}{\sqrt{2}}(|\uparrow_z\rangle_1 |\downarrow_{z'}\rangle_2 + |\downarrow_z\rangle_1 |\uparrow_{z'}\rangle_2)\\ |\psi_4\rangle &= \frac{1}{\sqrt{2}}(|\uparrow_z\rangle_1 |\downarrow_{z'}\rangle_2 - |\downarrow_z\rangle_1 |\uparrow_{z'}\rangle_2)\end{aligned} \tag{3b}$$

with spin polarized "up" or "down" along directions z and z'.

Operators of type (3a), although they refer to two separated spins, are effectively local. They can be measured simply by measuring the z component of spin of the first particle and the z' component of spin of the second particle. Operators with the eigenstates (3b) are truly nonlocal. They can be measured[7] in the same way as an operator with eigenstates given in Eq. (2) (a particular case of Eq. (3b)).

On the other hand[7] measurability of any nondegenerate operator with eigenstates not equivalent to the forms (3a) or (3b) implies the possibility of superluminal communication, i.e., violation of relativistic causality.

4. State Verification Measurements

A measurement of a nondegenerate operator is also a state verification measurement for all its eigenstates. The weakest possible definition of a state verification measurement which requires only *reliability* of the measurement is: the verification measurements of the state $|\psi_0\rangle$ must always yield the answer "yes" if the measured system has the initial state $|\psi_0\rangle$, and must always yield "no" if the system is initially in an orthogonal state. One may suspect that the verification of a state with canonical form (Schmidt decomposition) different from

$$\frac{1}{\sqrt{2}}(|\uparrow_z\rangle_1 |\uparrow_{z'}\rangle_2 + |\downarrow_z\rangle_1 |\downarrow_{z'}\rangle_2) \tag{4}$$

(the form of the eigenstates in (3b)) contradicts relativistic causality; i.e., that verification of a state

$$|\psi_1\rangle = \alpha |\uparrow_z\rangle_1 |\uparrow_{z'}\rangle_2 + \beta |\downarrow_z\rangle_1 |\downarrow_{z'}\rangle_2, \qquad |\alpha| \neq |\beta| \neq 0 \tag{5}$$

allows superluminal communication. Indeed, it has been shown[6] that the type of measurements of entangled states described above, i.e. nondemolition operator measurements with solely local interactions, cannot measure the state given by the form (5).

However, an unmeasurable quantity should not represent physical reality. If we want to consider the quantum state as a physical (versus purely mathematical) concept, it must be measurable. We do know how to ***prepare*** this state (the preparation procedure is also frequently called measurement). But the state (5) can also be measured using a new type of verification measurement named an ***exchange*** measurement[6] The idea is to make simultaneous short local interactions with parts of the measuring device such that the states of the system and the measuring device will be exchanged. The novel point in this method is that *local* interactions exchange *nonlocal* states. The result of the measurement cannot be read by two local observers; we must bring the two parts of the measuring device to one place. In addition, this procedure has another unconventional property. The final state of the system is completely independent of its initial state: it is just the initial state of the measuring device. The state of the system *is completely erased* by this state verification measurement.

It has recently been proven[7] that *any* verification of the state

$$|\psi_1\rangle = \alpha|\uparrow_z\rangle_1|\uparrow_{z'}\rangle_2 + \beta|\downarrow_z\rangle_1|\downarrow_{z'}\rangle_2, \qquad \alpha,\beta \neq 0 \tag{6}$$

erases all local information. The probable outcome of a local spin measurement performed after the state verification measurement is independent of the state of the composite system prior to the state verification. The example considered above of a measurable nondegenerate operator (2) trivially fulfills this result: for all eigenstates we have the property that the probability for any outcome of local spin measurement is the same. There is no local information after this nonlocal measurement.

5. Conclusions

Let us formulate the last result for the somewhat more general case of a system of two separated particles with several orthogonal states. Consider the Schmidt decomposition of a state $|\psi_0\rangle$ of this composite system:

$$|\psi_0\rangle = \sum_i \alpha_i |i\rangle_1 |i\rangle_2. \tag{7}$$

Here $|i\rangle_1$ and $|i\rangle_2$ are local orthonormal bases of states of the two particles. Let us denote by $H^{(1)}$ and $H^{(2)}$ the Hilbert spaces of part 1 and part 2 respectively, and by $H_0^{(1)}$ and $H_0^{(2)}$ the subspaces of $H^{(1)}$ and $H^{(2)}$ which are spanned by the base vectors $|i\rangle_1$ and $|i\rangle_2$ corresponding to coefficients $\alpha_i \neq 0$. Then for all initial states which belong to the Hilbert space $H_0^{(1)} \otimes H^{(2)}$, the probabilities $p(\psi)$ for results of local measurements in part 1, performed after verification of the state $|\psi_0\rangle$, have no dependence on the initial state.

Thus, the erasing effect of the proposed "exchange" measurements is a generic property of any reliable, causal state verification measurement. The full implications of this result are not yet clear. It already has helped complete the analysis of measurability of nondegenerate operators discussed above. It also has been used to show[7] that measurability of certain *ideal measurements of the first kind* contradicts relativistic causality, thus placing a serious doubt concerning the possibility of generalizing axiomatic quantum theory to the relativistic domain.

We would like to conclude by stressing the importance of measuring nonlocal properties via local interactions (with separate parts of the measuring device prepared in an entangled state). The same method can be used for so-called "multiple-time" measurements[8] which open the way to many new quantum phenomena.[9]

Acknowledgements

The research was supported by grant 425/91-1 of the the Basic Research Foundation (administered by the Israel Academy of Sciences and Humanities).

References

1. L. Landau and R. Peierls, *Z. Physik* **69** (1931) 56.
2. D. Bohm and Y. Aharonov, *Phys. Rev.* **108** (1957) 1070.
3. J. S. Bell, *Physics* 1 (64) 195.
4. Y. Aharonov and D. Albert, *Phys. Rev.* **D21** (1980) 3316.
5. Y. Aharonov and D. Albert, *Phys. Rev.* **D24** (1981) 359.
6. Y. Aharonov, D. Albert, and L. Vaidman, *Phys. Rev.* **D34** (1986) 1805.
7. S. Popescu and L. Vaidman, "Causality restrictions on nonlocal quantum measurements", preprint of Tel-Aviv University TAUP-2011-92 (1992).
8. Y. Aharonov and D. Albert, *Phys. Rev.* **D29** (1984) 223.
9. Y. Aharonov and L. Vaidman, *J.Phys. A: Math. Gen.* **24** (1991) 2315.

QUANTUM ANOMALIES AND THREE FAMILIES

PAUL H. FRAMPTON
Institute of Field Physics
Department of Physics and Astronomy
University of North Carolina
Chapel Hill, NC 27599-3255

ABSTRACT

Chiral anomalies and their cancellation are a fundamental quantum effect in relativistic field theory and can be fruitfully regarded as a topological phenomenon related to the Aharanov-Bohm effect. A possible relationship of such anomaly cancellation to the occurrence of three quark-lepton families is discussed.

It is a delightful honor to write for the sixtieth birthday of Yakir Aharanov with whom I have enjoyed many stimulating discussions about physics. Although best known for his contributions to the foundations of quantum mechanics, Yakir's broad knowledge makes him a useful colleague concerning any topic in theoretical physics.

In gauge field theory, chiral anomalies reflect a fundamental aspect of quantum theory and are a topological phenomenon related to the Aharanov-Bohm effect [1].

Gauge field theory is the basis of the successful standard model of particle interactions. In such a theory one first constructs a gauge-invariant lagrangian $L_B (\phi_B, \partial_\mu \phi_B)$ with bare quantities. At the quantum level, one wishes to renormalize to

$$L_B = L_R (\phi_R, \partial_\mu \phi_R) + \text{Counter-terms}$$

such that L_R is invariant under a gauge invariance isomorphic to the original one. This requires satisfaction of Taylor-Slavnov identities.

One peculiar Feynman diagram, a closed fermion loop with three gauge bosons attached (triangle diagram) spoils the possibility of such renormalization because of the chiral anomaly [2]. Unless this anomaly is cancelled by appropriate choice of chiral fermion representation of the gauge group the field theory is internally inconsistent and violates the requirements of renormalizability and unitarity.

The anomaly may be calculated locally through the Feynman diagram, or by global topological consideratıcns of the Atiyah-Singer index [3]. The second approach makes clear how the chiral anomaly [2] is a sequel to the AB effect [1].

Having established that connection, I now relate the AB effect further to the flavor question: why are there six flavors of quark u,d,s,c,b,t? It has long been thought that the replication of the quark flavors may be a result of anomaly concellation. For example, in 1979 there were attempts using SU(N) grand unification to find simple representations which led to three families under an SU(5) subgroup [4]; that program had some successes but did not really answer the basic question in a convincing manner.

In the standard model the chiral anomaly is cancelled between quarks and leptons in each family. This cancellation can be made to look non-trivial; e.g. for Y^3 the particles $(u, d)_L$ $\bar{u}_L$, $\bar{d}_L$, $(\nu, e)_L$ and $\bar{e}_L$ give the contributions

$$6(1/6)^3 + 3(-2/3)^3 + 3(1/3)^3 + 2(-1/2)^3 + (1)^3$$

which add to zero. Actually, this reflects the vanishing average electric charge. In any case, it gives no insight on why the flavor number equals six.

Our proposal [5] is to extend the standard electroweak gauge group to $SU(3)_L \times U(1)_X$ and to assign the leptons to antitriplets e.g. $(e^-, \nu_e, e^+)_L$ with $X = 0$ and similarly for the second and third families. The quarks of the first family are in the triplet $(u,d,D)_L$ where $Q(D) = 4/3$ and similarly $(c,s,S)_L$ for the second family. In the third family the quarks are assigned to an antitriplet $(b,t,T)_L$ with $Q(T) = +5/3$. The X values are respectively $-1/3$, $-1/3$, $+2/3$ leading to a cancellation of anomalies between the families each of which is separately anomalous.

To break the symmetry to $SU(2)_L \times U(1)_Y$ the Higgs sector contains a triplet with $X = +1$. All three exotic quarks acquire mass, as do five gauge bosons: the Z' and dileptons (Y^{--}, Y^-), (Y^{++}, Y^+). Because of $Z' - Z$ mixing the relevant scale is limited below by $M(Z') > 300$ GeV and $M(Y) > 230$ GeV, this last being coincident with the empirical lower limit [6]. At first sight, the new scale appears to be unrestricted from above but this is not the case for an interesting reason. The group theory of embedding $SU(2) \times U(1)$ in $SU(3) \times U(1)$ requires that $\sin^2\theta < 1/4$. Phenomenologically $\sin^2\theta = 0.233$ at $\mu = M_Z$ and increases with μ, becoming 0.25 at $\mu \approx 2.2$ TeV. This limit is singular, however, and g_X becomes strong-coupled so the upper limit is more nearly $M(Z') \leq 1$ TeV and both the Z' and Y are hence accessible to the supercollider.

In summary, the chiral anomaly of quantum theory dictates the chiral fermion content. Explication of flavor predicts dileptons (and Z' plus exotic quarks) at SSC.

Incidentally this "331" model gives insight into other features of the standard model from a new perspective, such as flavor-changing neutral currents and the GIM mechanism, charge quantization, possible neutrino masses and grand unification. These questions are under investigation.

This work was supported in part by the U.S. Department of Energy under Grant No. DE-FG05-95ER-40219.

References

1. Y. Aharanov and D. Bohm, *Phys. Rev.* **115**, 485 (1959)
2. S.L. Adler, *Phys. Rev.* **177**, 2426 (1969)
   J.S. Bell and R. Jackiw, *Nuov. Cim.* **60A**, 47 (1969)
3. M. Atiyah and I.M. Singer, *Publ. Math. Institut des Hautes Etudes Sci.* **27**, 305 (1969)
4. H. Georgi, *Nucl. Phy.* **B156**, 126 (1979)
   P.H. Frampton, *Phys. Lett.* **88B**, 299 (1979)
5. P.H. Frampton, *Phys. Rev. Lett.* **69**, 2889 (1992)
6. E.D. Carlson and P.H. Frampton, *Phys. Lett.* **B282**, 123 (1992)

EXPERIMENTS PURSUANT TO DETERMINING THE DURATION OF BARRIER TRAVESAL IN QUANTUM TUNNELING

M. J. Hagmann and L. Zhao
Department of Electrical and Computer Engineering, Florida International University
Miami, FL 33199, USA

ABSTRACT

Laser/STM experiments based on modulation of the barrier height by the electric field of light will be used to examine the duration of barrier traversal. The STM built for these measurements has decreased noise and improved stability. Our calculations suggest that a 670 nm laser diode at a power density of 100 W/cm^2 will reduce the tunneling current, which is contrary to most phenomena caused by laser illumination.

The question of tunneling times (i.e. traversal, reflection, and dwell times) has practical significance and has been the focus of much interest and controversy [1]. Measurements of tunnel conductance in heterostructures [2] and experiments with Josephson junctions [3] suggest that a specific time is associated with barrier traversal. Quantum mechanics provides useful results regarding tunneling but does not describe the motion of particles within the classically forbidden region.

A variety of theoretical procedures has been used to determine tunneling times, with different results [1]. Most of these methods give a definite value of traversal time for a specific problem, which appears inconsistent with the statistical nature of quantum phenomena. Distributions of tunneling times have been predicted using Feynman path integrals [4] and Bohm's causal interpretation of quantum mechanics [5], but they do not agree.

We model tunneling [6] on the basis of energy fluctuations consistent with the uncertainty principle. Cohen [7] postulated that the probability of a fluctuation decreases exponentially with the product of the magnitude ΔE and duration Δt, which product we refer to as the action of a fluctuation. He did not treat tunneling times, and considered only the most probable fluctuations (minimum action permitting tunneling), thus classically deriving the WKB solution for an opaque barrier. In previous work we modeled the full range of fluctuations to obtain distributions of the time for traversing static rectangular barriers. For large barriers these distributions are leptokurtic and centered at the semiclassical time (the classical time for traversing the inverted barrier [8]). The distributions are platykurtic (broad) for small barriers.

Several experimental methods have been used to examine the duration of tunneling. 1) Analyses of measured tunnel conductance [2] suggest that the response of image charges varies with barrier length. The location of the crossover from static to dynamic response appears consistent with the semiclassical traversal time, but there is much scatter in the data so this result is not definitive. 2) The effect of a magnetic field on tunnel conductance was studied [9], but the data may be explained without reference to tunneling times [10]. 3)Tunneling dynamics was studied in a shunted Josephson junction [3] but this involves tunneling between the states of a device and does not directly relate to tunneling by particles. 4) An operational tunneling time was determined from current rectification in a laser-illuminated scanning tunneling microscope (STM) [11], but the measured decrease in current with increasing barrier length may be explained without reference to tunneling times.

We have begun a project in which laser/STM experiments will be made with the objective of determining the duration of barrier traversal, but our work is based on barrier modulation rather than current rectification used in earlier studies [11]. Laser illumination of an STM junction modulates the barrier height. Theory suggests [8] that tunneling has two distinct regimes as a function of frequency, the crossover between them occurring when the angular frequency of the modulation equals the reciprocal of the traversal time.

We are completing a novel STM designed for these experiments. The circuit is similar to that of Park and Quate [12], but customized for decreased noise and increased stability. The preamplifier is chopper stabilized, and periodic multivalued illumination is used with boxcar signal averaging. Double sample and hold circuits minimize the droop while feedback is disengaged when the laser is pulsed. The quadrant electrodes on the piezoelectric tube scanner are fed with balanced X and Y supplies, and the inner electrode is fed with an unbalanced Z supply to provide orthogonal positioning of the tip. Power MOS-FET's are used in place of bipolar transistors in the high voltage sections to increase isolation and lower noise. A differential micrometer with a crossed-roller translation stage provides increased mechanical stability.

We have modeled the effect of a laser on the current in an STM. A rectangular barrier was assumed, but more appropriate expressions for the potential [13] will be implemented in later studies. We divide the barrier length into N segments of length d/N, such that each part is small enough that the potential is approximately a constant V during transit by an electron. We assume that an energy fiuctuation causes the particle to traverse each segment, and set $\Delta E = V - E + mv^2/2$, where the particle has mass m, velocity v, and nonperturbed energy E. In the present calculations, within each segment we consider only the most probable fluctuations, those with the least action permitting tunneling. Thus, within each segment the velocity $v = \sqrt{2(V-E)/m}$, the action of the fluctuation $A = d\sqrt{2m(V-E)}/N$, and the traversal time $T = d\sqrt{m/2(V-E)}/N$ which is the semiclassical value [6].

In each simulation the values of T and A are calculated within each segment, using the instantaneous value of the modulated potential for V, and summed to determine T, and A, which are the traversal time and action for tunneling through the entire barrier.This calculation is made for M different values of the modulation

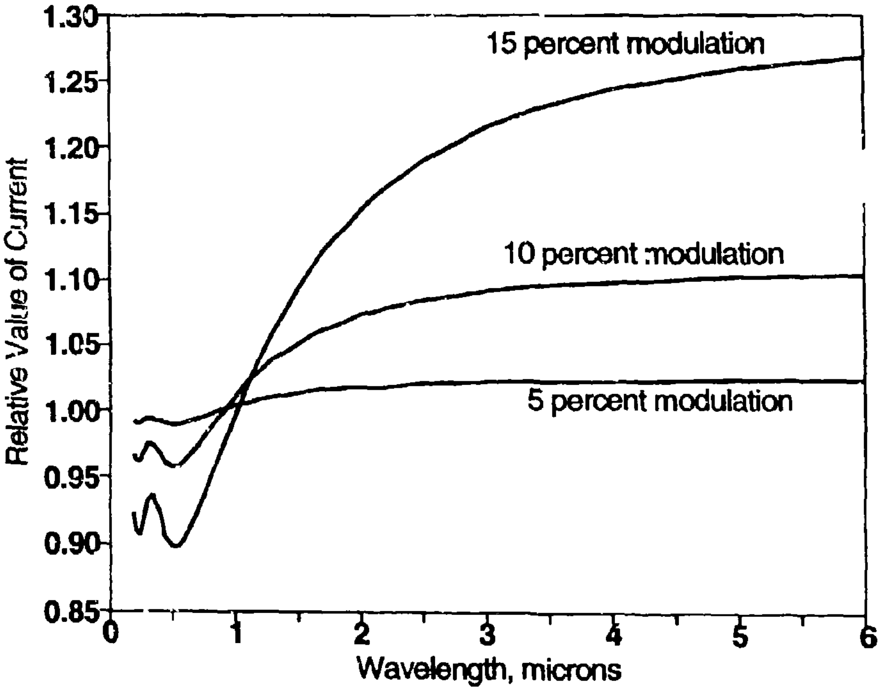

15 percent modulation
10 percent :modulation
5 percent modulation
Relative Value of Current
1.30
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0
1
2
3
4
5
6
Wavelength, microns

Fig. 1.

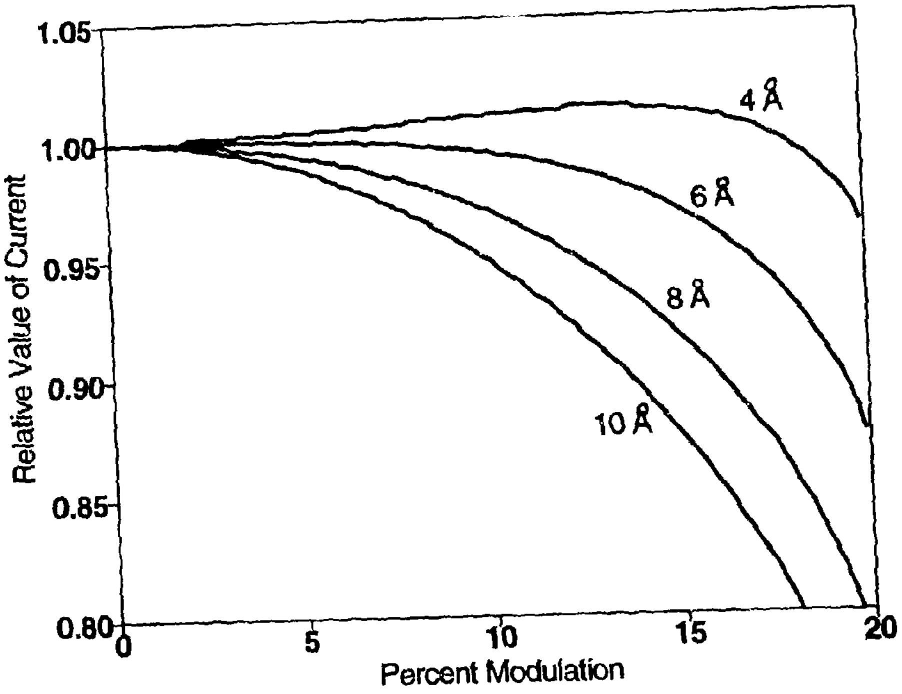

1.05
1.00
0.95
0.90
0.85
0.80
Relative Value of Current
0
5
10
15
20
Percent Modulation
4 Å
6 Å
8 Å
10 Å

Fig. 2.

phase at which the electron enters the barrier. For each entry phase the transmission is determined by assuming the probability of the fluctuation is proportional to $\exp(-A_s/\hbar)$. Finally, the mean value is normalized by dividing by the value without modulation, to obtain the relative value of the current with modulation. The values of N and M are increased untill these parameters are found to have negligible effect. The results of several simulations are presented in the following two figures.

Figure 1 shows the relative current as a function of the modulating-wavelength for 4.0 eV electrons with a barrier length of 6.0 Å and height of 5.0 V. The three levels of modulation correspond to illumunation at different power densities. Since there is a distribution of traversal times, instead of the definite values of time implicit in other analyses [8], the transition between the regime for low and high modulation frequencies is broad. Our long-term objective is to examine this crossover by measuring the current when two or more lasers sequentially illumine an STM junction with similar power densities at different frequencies.

If the modulation frequency is high enough, Fig.1 shows that barrier modulation tends to inhibit tunneling. Figure 2 shows the relative current as a function of the level of modulation for 4.0 eV electrons with a barrier height of 5.0 V and a modulating wavelength of 670 nm. The data in Fig.2 suggest that for barrier lengths between 6 and 10 Å the current decreases as the power density is increased. In our first experiments we will determine the effects of power density on tunneling current when the STM junction is illuminated with a 670 nm laser diode. A power density of 100 W/cm^2 , providing adequate modulation, may be obatined with a 20 mW laser diode focused to a minimal spot size.

A variety of phenomena occur in laser/STM experiments [14] including current rectification, photo-assisted tunneling, thermal-assisted tunneling and thermal expansion, as well as the effects which we have modeled. In our initial experiments the use of a visible laser would decrease current rectification [15]. The relatively low power density would reduce thermal effects as well as current rectification, but all of these effects must be considered. Subsequent experiments made to examine the crossover in Fig.1 would be more difficult to interpret because the competing phenomena have different frequency dependence.

References

1. For reviews, see A. P, Jauho, *Hot Carriers in Semiconductor Nanostructures*, edited by J. Shah (Academic Press, New York, 1992), pp. 121-151; M. Büttiker, in *Electronic Properties of Multilayers and Low-Dimensional Semiconductor Structure*, edited by J. M. Chamberlain, L. Eaves and J. -C. Portal (Plenum, New York, 1990), pp. 297-315.
2. P. Gueret, E. Marclay and H. Meier, *Appl. Phys. Lett.* 53, 1617 (1988).
3. D. Esteve, J. M. Martinis, C. Urbina, E. Turlot M. H. Devoret, H. Grabert and S. Linkwitz, *Physica Scr.* T29, 121 (1989).
4 H. A. Fertig, *Phys.Rev.Lett.* 65, 2321 (1990). D. Sokolovski and J. N. L.

Connor, *Phys. Rev.* A44, 1500 (1991).

5. C. R. Leavens, *Solid State Commun.* 76, 253 (1990).
6. M. J. Hagmann, *Solid State Commun.* 82, 867 (1992).
7. B. L. Cohen, *Am. J. Phys.* 33, 97 (1965).
8. M. Büttiker and R. Landauer, *Phys. Rev. Lett.* 49, 1739 (1982). M. Büttiker and R.Landauer, *Physica Scr.* 32, 429 (1985).
9. P. Gueret, A. Baratoff and E. Marclay, *Europhys. Lett.* 3, 367 (1987).
10. L. Eaves, K. W. H. Stevens and F. W. Sheard, in *The Physics and Fabrication of Microstructures and Microdevices*, edited by M. J. Kelly and C. Weisbuch (Springer, Berlin, 1986).
11. H. Q. Nguyen, P. H. Cutler, T. E. Feuchtwang, Z. H. Huang, Y. Kuk, P. J. Silverman, A. A. Lucas, and T. E. Sullivan, *IEEE Trans. Electron Devices* 36, 2671 (1989); A. A. Lucas, P. H. Cutler, T. E. Feuchtwang, T. T. Tsong, T. E. Sullivan, Y. Kuk, H. Nguyen and P. J. Silverman, *J. Vac. Sci. Technol.* A 6, 461 (1988).
12. S. Park and C. F. Quate, *Rev. Sci. Inst.* 58, 2010 (1987).
13. N. M. Miskovski, P. H. Cutler, T. E. Feuchtwang and A. A. Lucas,*Int J. Infrared Millim. Waves* 2, 739 (1981); L. E. Bar'yudin, V. L. Bulatov and D. A. Telnov, *J. App. Phys.* 71, 946 (1992).
14. Z. Hassan, D. Andsager, D. Saltz, K. Cartwright and M. H. Nayfeh, *Rev. Sci. Inst.* 63, 2099 (1992).
15. A. Sanchez, C. F. Davis Jr., K. C. Liu and A. Javan, *J. Appl. Phys.* 49, 5270 (1978).

Figure Captions

Fig. 1. Relative current vs. modulating-wavelength for 4.0 eV electrons with barrier length = 6.0 Å and height = 5.0 V.

Fig. 2. Relative current vs. level of modulation for 4.0 eV electrons with barrier height = 5.0 V and modulating wavelength = 670 nm.

GLOBAL QUANTIZATION OF VACUUM ANGLE AND MAGNETIC MONOPOLES AS A NEW SOLUTION TO THE STRONG CP PROBLEM

HUAZHONG ZHANG
Department of Physics and Atmospheric Sciences, Jackson State University,
P.O.Box 17660, Jackson, MS 39217, USA[1]
and
Theoretical Physics Group, Lawrence Berkeley Laboratory
Berkeley, CA 94720, USA[2]

ABSTRACT

The non-perturbative solution to the strong CP problem with magnetic monopoles as originally proposed by the author is described. It is shown that the gauge orbit space with gauge potentials and gauge tranformations restricted on the space boundary and the globally well-defined gauge subgroup in gauge theories with a θ term has a monopole structure if there is a magnetic monopole in the ordinary space. The Dirac's quantization condition then ensures that the vacuum angle θ in the gauge theories must be quantized to have a well-defined physical wave functional. The quantization rule for θ is derived as $\theta = 0, 2\pi/n$ $(n \neq 0)$ with n being the topological charge of the magnetic monopole. Therefore, the strong CP problem is automatically solved with the existence of a magnetic monopole of charge ±1 with $\theta = \pm 2\pi$. This is also true when the total magnetic charge of monopoles are very large ($|n| \geq 10^{9} 2\pi$). The fact that the strong CP violation can be only so small or vanishing may be a signal for the existence of magnetic monopoles and the universe is open.

1. Introduction and Summary of the Main Results

Yang-Mills theories[1] and their non-perturbative effects have played one of the most important roles in particle physics. It is known that, in non-abelian gauge theories a Pontryagin or θ term,

$$\mathcal{L}_\theta = \frac{\theta}{32\pi^2} \epsilon^{\mu\nu\lambda\sigma} F^a_{\mu\nu} F^a_{\lambda\sigma}, \tag{1}$$

can be added to the Lagrangian density of the system due to instanton[2] effects in gauge theories. This term can induce CP violations for an abitrary value of θ. Especially, such an effective θ term in QCD may induce CP violations in strong interactions. In our discussions relevant to QCD, θ is simply used to denote $\theta + arg(detM)$ effectively with M being the quark mass matrix, when the effects of electroweak interactions are included. However, the experimental results on the neutron electric

[1] Permanent address

[2] Where the work supported by a science consortium award of DOE.

dipole moment strongly limit the possible values of the θ in QCD ($\leq 10^{-9}$, modulo 2π for example). This is the well-known strong CP problem. One of the most interesting understanding of the strong CP problem has been the assumption of an additional Peccei-Quinn $U(1)_{PQ}$ symmetry[4], but the observation has not given[3] evidence for the axions[5] needed in this approach. Thus the other possible solutions to this problem are of fundamental interest.

Recently, a non-perturbative solution to the Strong CP problem with magnetic monopoles has been proposed originally by the author[6]. In our solution[6], it is proposed that the vacuum angle with magnetic monopoles must be quantized. Our quantization rule is derived essentially by two different methods. This is given by $\theta = 0$, or $\theta = 2\pi N/n$ ($n \neq 0$) with integer n being the relevant topological charge of the magnetic monopole and N may be fixed as 1 in the method 1 and is an arbitrary integer in the method 2. The first method[1] is to show the existence of a monopole structure in the relevant gauge orbit space in Scherodinger formulation[7,8], and using the Dirac quantization rule for having a well-defined wave funtional. The second method is to show that there exist well-defined gauge transformations which will ensure the quantization of θ by the constraints of Gauss's law due to the non-abelican electric charges carried by the magnetic monopoles proportional to θ as noted in Ref. 21 and generalized in Ref.22 to the non-abelian case for the generalized magnetic monopoles[17].

Therefore, we conclude that strong CP problem can be solved due to the quantization of θ in the presence of magnetic monopoles, for example monopoles of topological charge $n = \pm 1$ with $\theta = \pm 2\pi$, or $n \geq 2\pi 10^9$ with $\theta \leq 10^{-9}$. Moreover, the existence of non-vanishing magnetic flux through the space boundary implies that the universe must be open. In this note, we will briefly describe and review our solution to the strong CP problem with magnetic monopoles with the first method.

2. Quantization Condition on θ and Solution to the Strong CP Problem

The main idea of our discussions is based on the follows. A wave functional in the gauge orbit space corresponds to a cross section[12] of the relevant fiber bundle for the theory. Topologically, if there is a non-vanishing gauge field as the curvature in the gauge orbit space, then the flux of the curvature through a closed surface in the gauge orbit space must be quantized to have a cross section[7-8]. Physically, this is equivalent to say that the magnetic flux through the closed surface must be quantized according to the Dirac quantization condition in order to have a well-defined wave functional in the quantum theory.

In this method, we will extend the method of Wu and Zee in Ref.7 for the discussions of the effects of the Pontryagin term in pure Yang-Mills theories in the gauge orbit spaces in the Schrodinger formulation. This formalism has also been used with different methods to derive the mass parameter quantization in three-dimensional Yang-Mills theory with Chern-Simons term[7-8]. It is shown in Ref. 7 that the Pontryagin term induces an abelian background field or an abelian structure in the gauge configuration space of the Yang-Mills theory. In our discussions, we will consider the case with the existence of a magnetic monopole. We will show that magnetic monopoles[9-10] in space will induce an abelian gauge field with non-vanishing field strength in gauge configuration space, and magnetic flux through a two-dimensional sphere in the induced gauge orbit space on the space boundary is non-vanishing. Then, Dirac condition[9-10] in the corresponding quantum theories leads to the result that the relevant vacuum angle θ must be quantized as $\theta = 2\pi/n$ with n being the topological charge of the monopole to be generally defined. Therefore, the strong CP problem can be solved with the existence of magnetic monopoles.

We will now consider the Yang-Mills theory with the existence of a magnetic monopole at the origin. Our derivation applies generally to a gauge theory with an arbitrary simple gauge group or a U(1) group outside the monopole. This gauge group under consideration may be regarded as a factor group in the exact gauge group of a grand unification theory. Note that there can be Higgs field and unification gauge fields confined inside the monopole core, which will be ignored in our discussion outside the monopoles.

As we will see that an interesting feature in our derivation is that we will use the Dirac quantization condition both in the ordinary space and restricted gauge orbit space to be defined. The Lagrangian of the system is given by

$$\mathcal{L} = \int d^4x \{ -\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} + \frac{\theta}{32\pi^2} \epsilon^{\mu\nu\lambda\sigma} F^a_{\mu\nu} F^{a\lambda\sigma} \}. \tag{2}$$

We will use the Schrodinger formulation and the Weyl gauge $A^0 = 0$. The conjugate momentum corresponding to A^a_i is given by

$$\pi^a_i = \frac{\delta \mathcal{L}}{\delta \dot{A}^a_i} = \dot{A}^a_i + \frac{\theta}{8\pi^2} \epsilon_{ijk} F^a_{jk}. \tag{3}$$

In the Schrodinger formulation, the system is similar to the quantum system of a particle with the coordinate q_i moving in a gauge field $A_i(q)$ with the correspondence[6-

$$q_i(t) \rightarrow A^a_i(\mathbf{x}, t), \tag{4}$$

$$A_i(q) \rightarrow \mathcal{A}^a_i(\mathbf{A}(\mathbf{x})), \tag{5}$$

where

$$\mathcal{A}^a_i(\mathbf{A}(\mathbf{x})) = \frac{\theta}{8\pi^2} \epsilon_{ijk} F^a_{jk}. \tag{6}$$

Thus there is a gauge structure with gauge potential $\mathcal{A}$ in this formalism within a gauge theory with the θ term included. Note that in our discussion with the presence of a magnetic monopole, the gauge potential $\mathbf{A}$ outside the monopole generally need to be understood as well defined in each local coordinate region. In the overlapping regions, the separate gauge potentials can only differ by a well-defined gauge transformation[10]. In fact, single-valuedness of the gauge function corresponds to the Dirac quantization condition[10]. For a given r, we can choose two extended semi-spheres around the monopole, with $\theta \in [\pi/2 - \delta, \pi/2 + \delta](0 < \delta < \pi/2)$ in the overlapping region, where the θ denotes the θ angle in the spherical polar coordinates. For convenience, we will use differential forms[10] in our discussions, where $A = A_i dx^i, F = \frac{1}{2} F_{jk} dx^j dx^k$, with $F = dA + A^2$ locally. For our purpose to discuss about the effects of the abelian gauge structure on the quantization of the vacuum angle, we will now briefly clarify the relevant topological results needed, then we will realize the topological results explicitly.

With magnetic monopoles, we need to generalize the gauge orbit space of ordinary gauge theories to include the space boundary which is noncontractible with non-vanishing magnetic flux quantized according Dirac quantization condition. With the constraint of Gauss' law, the quantum theory in the finite space region in this formalism is described in the usual gauge orbit space $\mathcal{U}/\mathcal{G}$. The $\mathcal{U}$ is the space of well-defined gauge potentials and $\mathcal{G}$ denotes the space of continuous gauge transformations with gauge functions mapping the space boundary to a single point in the gauge group. Due to the exitence of magnetic monopoles, the gauge transformations on the space boundary S^2 can be non-trivial, the physical effects of the well-defined gauge transformations need to be considered. As it is known that[12], only the gauge transformations generated by the generators commuting with magnetic charges may be well-defined globally. On the space boundary, $\mathcal{U}$ will also be used to denote the induced gauge configuration space with gauge potentials restricted on the space boundary, and $\mathcal{G}$ also denotes the continuous gauge transformations restricted on the space boundary and well-defined gauge subgroup. Then we will call corresponding $\mathcal{U}/\mathcal{G}$ as restricted gauge orbit space. Collectively, they will be called as the usual space for the finite coordinate space region and the restricted space on the space boundary. There should not be confusing for the notations used both for the usual spaces and restricted spaces. As we will see that the magnetic charges up to a conjugate transformation are in a Cartan subalgebra of the gauge group, then on the space boundary S^2, we need to consider a well-defined gauge subgroup $G = U(1)$ for the quantization of θ. Similar to the usual gauge orbit space on the compactified coordinate space by restricting to gauge functions mapping the space boundary to a single point in the guage group, the restricted gauge orbit space is well-defined since the space boundary S^2 is compact.

Note that the physical meaning of the restricted gauge orbit space can be understood as follows. Let $\Psi_{phys}(\mathbf{A}(\mathbf{x}))$ denote the physical wave functional and $\Psi_{phys}(\mathbf{A}(\mathbf{x})) |_{S^2}$ be its restriction on the space boundary S^2 which actually only depends on the direction of $\mathbf{x}$. Then, the $\Psi_{phys} |_{S^2}$ must be invariant under the gauge transformations well-defined on the entire space boundary. Namely the $\Psi_{phys} |_{S^2}$ is defined in the restricted gauge orbit space. However, in the finite space region, the $\Psi_{phys}(\mathbf{A}(\mathbf{x}))$ for finite $\mathbf{x}$ is only required to be invariant under the gauge transformations with gauge function going to the identity at the spatial infinity. Namely, it is defined the usual gauge orbit space. The entire Ψ_{phys} is then well-defined in the generalized gauge orbit space as described.

Now consider the following exact homotopy sequence[13] both for the ususal and restricted spaces:

$$\Pi_N(\mathcal{U}) \xrightarrow{P_*} \Pi_N(\mathcal{U}/\mathcal{G}) \xrightarrow{\Delta_*} \Pi_{N-1}(\mathcal{G}) \xrightarrow{i_*} \Pi_{N-1}(\mathcal{U}) \ (N \geq 1). \tag{7}$$

Note that homotopy theory has also been used to study the global gauge anomalies [14–16], especially by using extensively the exact homotopy sequences and in terms of James numbers of Stiefel manifolds[17]. One can easily see that $\mathcal{U}$ is topologically trivial, thus $\Pi_N(\mathcal{U}) = 0$ for any N. Since the interpolation between any two gauge potentials A_1 and A_2

$$A_t = tA_1 + (1-t)A_2 \tag{8}$$

for any real t is in $\mathcal{U}$ (Theorem 7 in Ref.10, and Ref.7). since A_t is transformed as a gauge potential in each local coordinate region, and in an overlapping region, both A_1 and A_2 are gauge potentials may be defined up to a gauge transformation, then A_t is a gauge potential which may be defined up to a gauge transformation, namely, $A_t \in \mathcal{U}$. Thus, we have

$$0 \xrightarrow{P_*} \Pi_N(\mathcal{U}/\mathcal{G}) \xrightarrow{\Delta_*} \Pi_{N-1}(\mathcal{G}) \xrightarrow{i_*} 0 \ (N \geq 1). \tag{9}$$

This implies that

$$\Pi_N(\mathcal{U}/\mathcal{G}) \cong \Pi_{N-1}(\mathcal{G}) \ (N \geq 1). \tag{10}$$

As we will show that in the presence of a magnetic monopole, the topological properties of the system are drastically different. This will give important consequences in the quantum theory. In fact, the topological properties of the restricted gauge orbit spaces are relevant for our purpose since as we will see that only the integrals on the sp[illegible] boundary S^2 are relevant in the quantization equation for the θ. Now for the re[illegible] spaces, the main topological result we will use is given by

$$\Pi_2(\mathcal{U}/\mathcal{G}) \cong \Pi_1(\mathcal{G}) = \Pi_1(G) \oplus \Pi_3(G), \tag{11}$$

for a well-defined gauge subgroup G. As we will see that in the relevant case of $G = U(1)$ for our purpose $\Pi_3(G) = 0$. The condition $\Pi_2(\mathcal{U}/\mathcal{G}) \neq 0$ corresponds to the existence of a magnetic monopole in the restricted gauge orbit space. We will first show that in this case $\mathcal{F} \neq 0$, and then demonstrate explicitly that the magnetic flux $\int_{S^2} \hat{\mathcal{F}} \neq 0$ can be nonvanishing in the restricted gauge orbit space, where $\hat{\mathcal{F}}$ denotes the projection of $\mathcal{F}$ into the restricted gauge orbit space.

Denote the differentiation with respect to space variable $\mathbf{x}$ by d, and the differentiation with respect to parameters $\{t_i \mid i = 1, 2...\}$ which $\mathbf{A}(\mathbf{x})$ may depend on in the gauge configuration space by δ, and assume $d\delta + \delta d = 0$. Then, similar to $A = A_\mu dx^\mu$ with μ replaced by a, i, $\mathbf{x}$, $A = A_i^a L^a dx^i, F = \frac{1}{2} F_{jk}^a L^a dx^j dx^k$ and $tr(L^a L^b) = -\frac{1}{2}\delta^{ab}$ for a basis $\{L^a \mid a = 1, 2, ..., rank(G)\}$ of the Lie algebra of the gauge group G, the gauge potential in the gauge configuration space is given by

$$\mathcal{A} = \int d^3x \mathcal{A}_i^a(\mathbf{A}(\mathbf{x}))\delta A_i^a(\mathbf{x}). \tag{12}$$

Using Eq.(6), this gives

$$\mathcal{A} = \frac{\theta}{8\pi^2} \int d^3x \epsilon_{ijk} F_{jk}^a(\mathbf{x}) \delta A_i^a(\mathbf{x}) = -\frac{\theta}{2\pi^2} \int_M tr(\delta A F), \tag{13}$$

with M being the space manifold. With $\delta F = -D_A(\delta A) = -\{d(\delta A) + A\delta A - \delta AA\}$, we have topologically

$$\mathcal{F} = \delta\mathcal{A} = \frac{\theta}{2\pi^2}\int_M tr[\delta A D_A(\delta A)] = \frac{\theta}{4\pi^2}\int_M dtr(\delta A\delta A) = \frac{\theta}{4\pi^2}\int_{\partial M} tr(\delta A\delta A). \qquad (14)$$

Usually, one may assume $A \to 0$ faster than 1/r as $\mathbf{x} \to \mathbf{0}$, then[7] this would give $\mathcal{F} = 0$. However, this is not the case in the presence of a magnetic monopole. Asymptotically, a monopole may generally give a field strength of the form[9-10,17]

$$F_{ij} = \frac{1}{4\pi r^2}\epsilon_{ijk}(\hat{\mathbf{r}})_k G(\hat{\mathbf{r}}), \qquad (15)$$

with $\hat{\mathbf{r}}$ being the unit vector for $\mathbf{r}$, and this gives $A \to O(1/r)$ as $\mathbf{x} \to 0$. Thus, one can see easily that a magnetic monopole can give a nonvanishing field strength $\mathcal{F}$ in the gauge configuration space. To evaluate the $\mathcal{F}$, one needs to specify the space boundary ∂M in the presence of a magnetic monopole. we now consider the case that the magnetic monopole does not generate a singularity in the space. In fact, this is so when monopoles appear as a smooth solution of a spontaneously broken gauge theory similar to 't Hooft Polyakov monopole[9]. For example, it is known that[18] there are monopole solutions in the minimal SU(5) model. Then, the space boundary may be regarded as a large 2-sphere S^2 at spatial infinity. For our purpose, we actually only need to evaluate the projection of $\mathcal{F}$ into the gauge orbit space.

In the gauge orbit space, a gauge potential can be written in the form of

$$A = g^{-1}ag + g^{-1}dg, \qquad (16)$$

for an element a $\in \mathcal{U}/\mathcal{G}$ and a gauge function $g \in \mathcal{G}$. Then the projection of a form into the gauge orbit space contains only terms proportional to $(\delta a)^n$ for integers n. We can now write

$$\delta A = g^{-1}[\delta a - D_a(\delta g g^{-1})]g. \qquad (17)$$

Then we obtain

$$\mathcal{A} = -\frac{\theta}{2\pi^3}\int_M tr(f\delta a) + \frac{\theta}{2\pi^2}\int_M tr[f D_a(\delta g g^{-1})], \qquad (18)$$

where $f = da + a^2$. With some calculations, this can be simplified as

$$\mathcal{A} = \hat{\mathcal{A}} + \frac{\theta}{2\pi^2}\int_{S^2} tr[f\delta g g^{-1}], \qquad (19)$$

where

$$\hat{\mathcal{A}} = -\frac{\theta}{2\pi^2}\int_M tr(f\delta a), \qquad (20)$$

is the projection of $\mathcal{A}$ into the gauge orbit space. Similarly, we have

$$\mathcal{F} = \frac{\theta}{4\pi^2}\int_{S^2} tr\{[\delta a - D_a(\delta g g^{-1})][\delta a - D_a(\delta g g^{-1})]\} \qquad (21)$$

or

$$\mathcal{F} = \hat{\mathcal{F}} - \frac{\theta}{4\pi^2}\int_{S^2} tr\{\delta a D_a(\delta g g^{-1}) + D_a(\delta g g^{-1})\delta a - D_a(\delta g g^{-1})D_a(\delta g g^{-1})\}, \qquad (22)$$

where

$$\hat{\mathcal{F}} = \frac{\theta}{4\pi^2}\int_{S^2} tr(\xi a\delta a). \tag{23}$$

Now ali our discussions will be based on the restricted spaces. To see that the flux of $\hat{\mathcal{F}}$ through a closed surface in the restricted gauge orbit space $\mathcal{U}/\mathcal{G}$ can be nonzero, we will construct a 2-sphere in it. Consider an element $a \in \mathcal{U}/\mathcal{G}$, and a loop in $\mathcal{G}$. The set of all the gauge potentials obtained by all the gauge transformations on a with gauge functions on the loop then forms a loop C^1 in the gauge configurations space $\mathcal{U}$. Obviously, the a is the projection of the loop C^1 into $\mathcal{U}/\mathcal{G}$. Now since $\Pi_1(\mathcal{U}) = 0$ is trivial, the loop C^1 can be continuously extented to a two-dimensional disc D^2 in the $\mathcal{U}$ with $\partial D^2 = C^1$, then obviously, the projection of the D^2 into the gauge orbit space is topologically a 2-sphere $S^2 \subset \mathcal{U}/\mathcal{G}$. With the Stokes' theorem in the gauge configuration space, We now have

$$\int_{D^2} \mathcal{F} = \int_{D^2} \delta\mathcal{A} = \int_{C^1} \mathcal{A}. \tag{24}$$

Using Eqs.(19) and (24) with $\delta a = 0$ on C^1, this gives

$$\int_{C^1} \mathcal{A} = \frac{\theta}{2\pi^2} tr \int_{S^2}\int_{C^1} [f\delta g g^{-1}]. \tag{25}$$

Thus, the projection of the Eq.(26) to the gauge orbit space gives

$$\int_{S^2} \hat{\mathcal{F}} = \frac{\theta}{2\pi^2} tr \int_{S^2} \{f \int_{C^1} \delta g g^{-1}\}, \tag{26}$$

where note that in the two S^2 are in the gauge orbit space and the ordinary space respectively. We have also obtained this by verifying that

$$\int_{D^2} tr \int_{S^2} tr\{\delta a D_a(\delta g g^{-1}) + D_a(\delta g g^{-1})\delta a - D_a(\delta g g^{-1}) D_a(\delta g g^{-1})\} = 0, \tag{27}$$

or the projection of $\int_{D^2} \mathcal{F}$ gives $\int_{S^2} \hat{\mathcal{F}}$.

In quantum theory, Eq.(26) corresponds to the topological result $\Pi_2(\mathcal{U}/\mathcal{G}) \simeq \Pi_1(\mathcal{G})$ on the restricted spaces. The discussion about the Hamiltonian equation in the schrodinger formulation will be similar to that in Refs.7 and 8 including the discussions for the three-dimensional Yang-Mills theories with a Chern-Simons term. We need the Dirac quantization condition to have a well-defined wave functional in the formalism. In the gauge orbit space, the Dirac quantization condition gives

$$\int_{S^2} \hat{\mathcal{F}} = 2\pi k, \tag{28}$$

with k being integers. The Dirac quantization condition in the gauge orbit space will be clarified shortly. Now let f be the field strength 2-form for the magnetic monopole. The quantization condition is now given by[17]

$$exp\{\int_{S^2} f\} = exp\{G_0\} = exp\{4\pi \sum_{i=1}^{r} \beta^i H_i\} \in Z. \tag{29}$$

Where G_0 is the magnetic charge up to a conjugate transformation by a group element, H_i (i=1, 2,...,r=rank(G)) form a basis for the Cartan subalgebra of the gauge group with simple roots α_i (i=1,2,...,r). We need non-zero topological value to obtain quantization condition for θ. As it is known from Ref.12, only the gauge transformations commuting with the magnetic charges can be globally well-defined, only those gauge transformations can be used for determining the global topological quantities. Consider g(x,t) in the well- defined U(1) gauge subgroup commutative with the magnetic charges on the C^1

$$g(x,t)\,|_{x\in S^2}= exp\{4\pi mt\sum_{i,j}\frac{(\gamma_i)^j H_j}{<\alpha_i,\alpha_i>}\}, \qquad (30)$$

with m being integers and $t\in[0,1]$. In fact, m should be identical to k according to our topological result $\Pi_2(\mathcal{U}/\mathcal{G})\cong\Pi_1(\mathcal{G})$. The k and m are the topological numbers on each side. Thus, we obtain in the case of non-vanishing vacuum angle θ

$$\theta=\frac{2\pi}{n}\ (n\neq 0). \qquad (31)$$

Where we define generally the topological charge of the magnetic monopole as

$$n=-2<\delta,\beta'> \qquad (32)$$

which must be an integer[17]. Where

$$\delta'=\sum_i\frac{2\alpha_i}{<\alpha_i,\alpha_i>}, \qquad (33)$$

the minus sign is due to our normalization convention for Lie algebra generators. Note that the parameter t of g(x,t) in eq.(30) may be regarded as the time parameter topologically when the time evolution is included, the two end points of the closed loop then correspond to the time infinities. The g(x,t) is not a constant in the entire spacetime, and does not generate a Nother symmetry. The non-trivial topological properties ensure that the non-trivial spacetime dependence will be maintained when continuous Lorentz transformations are implemented. Consequently, the requirement of gauge invariance corresponding to the g(x,t) will not eliminate any charged configurations.

Therefore, we conclude that in the presence of magnetic monopoles with topological charge ±1, the vacuum angle of non-abelian gauge theories must be $\pm 2\pi$, the existence of such magnetic monopoles gives a solution to the strong CP problem. But CP cannot be exactly conserved in this case since $\theta=\pm 2\pi$ correspond to two different monopole sectors. The existence of many monopoles can ensure $\theta\to 0$, and the strong CP problem may also be solved. In this possible solution to the strong CP problem with $\theta\leq 10^{-9}$, the total magnetic charges present are $|n|\geq 2\pi 10^9$. This may possibly be within the abundance allowed by the ratio of monopoles to the entropy[19], but with the possible existence of both monopoles and anti-monopoles, the total number of magnetic monopoles may be larger than the total magnetic charges. Generally, one needs to ensure that the total number is consistent with the experimental results on the abundance of monopoles. The $n = \pm 2$ may also possibilely solve CP if it is consistent with the experimental observation.

Note that we only considered non-singular magnetic monopole in the space. For 't Hooft Polykov monopole, the full gauge group inside the monopole is simply connected, it will not give any boundary contribution to the term in Eq.(26). However, outside the monopole, the gauge symmetry is spontaneously broken, it is known that the unbroken gauge group cannot be simply connected to have monopole solutions. For example, in SU(5) model, inside the monopole, SU(5) is simply connected; outside the monopole the exact gauge group G=SU(3)xU(1) satisfies $\Pi_1(G) = Z$. We expect that in general, the GUT monopoles are smooth solutions, and therefore cannot have a mathematical boundary at a given short distance around the monopole relevant to our boundary contribution. Therefore, the realistic world meet the condition to have our solution to the strong CP problem.

The effect of a term proportional to $\epsilon^{\mu\nu\lambda\sigma}F_{\mu\nu}F_{\lambda\sigma}$ in the presence of magnetic charges was first considered[20] relevant to chiral symmetry. The effect of a similar U(1) θ term was discussed for the purpose of considering the induced electric charges[21] as quantum excitations of dyons associated with the 't Hooft Polyakov monopole and generalized magnetic monopoles[16,21]. Note that since our solution needs non- vanishing magnetic flux through the space boundary, this implies that only an open universe car be consistent with our solution. Note that the relevance to the $U_A(1)$ problem is discussed in Ref. 23.

Acknowledgements

The author would like to express his gratitude to Y. S. Wu and A. Zee for valuable discussions. He is also grateful to O. Alvarez and the Theoretical Physics Group at the Lawrence Berkeley Laboratory for their hospitality during the major completion of the work.

References

1. C. N. Yang and R. L. Mills, *Phys. Rev.* **96** (1982) 191.
2. A. A. Belavin, A. M. Polyakov, A. S. Schwarz and Yu. Tyupkin, *Phys. Lett.* **B59** (1975) 85; R. Jackiw and C. Rebbi, *Phys. Rev. Lett.* **37** (1976) 172; C. Callan, R. Dashen and D. Gross, *Phys. Lett.* **B63** (1976) 334; G. 't Hooft,*Phys. Rev. Lett.* **37** (1976) 8; and *Phys. Rev.* **D14** (1976) 334.
3. R. Peccei, in *CP Violations*, ed. C. Jarlskog (World Scientific, Singapore, 1989).
4. R. D. Peccei and H. R. Quinn, *Phys. Rev. Lett.* **38** (1977) 1440; *Phys. Rev.* **D16** (1977) 1791.
5. S. Weinberg, *Phys. Rev. Lett.* **40** (1978) 223; F. Wilczek, *Phys. Rev. Lett.* 40 (1978) 271.

6. H. Zhang, *Preprint* **LBL-32491** June 1992; **LBL-32595** July 1992; and in *Proc. Division. Meeting of Particles and Fields of APS 1992 at Fermilab*, (World Scientific, Singapore).
7. Y. S. Wu and A. Zee, *Nucl. Phys.* **B258** (1985) 157.
8. R. Jackiw, in *Relativity, Groups and Topology II*, ed. B. DeWitt and R. Stora (Noth-Holland, 1984); *Proc. Argonne Anomaly Conference 1985.*
9. P. A. M. Dirac, *Proc. Roy. Soc.* **A133** (1931) 60. *Phys. Rev.* **74** (1948) 817; 't Hooft, *Nucl. Phys.* **79** (1974) 276; A. M. Polyakov, *JETP Lett.* **20** (1974) 194; or see S. Coleman, in *The Unity of the Fundamental Interactions.*
10. T. T. Wu and C. N. Yang, *Phys. Rev.* **D12** (1975) 3845.
11. See for example, B. Zumino, in *Relativity, Groups and Topology II, (Les Houches 1983)*, ed. B. DeWitt and R. Stora (Noth-Holland, 1984).
12. P. Nelson and A. Manohar, *Phys. Rev. Lett.* **50** (1983) 943; A. Balachandran, G. Marmo, M. Mukuanda, J. Nilsson, E. Sudarshan, and F. Zaccaria, *Phys. Rev. Lett.***50** (1983) 1553; A. Abouelsaood, *Phys. Lett.* **B125**, (1983) 467.
13. S. T. Hu, *Homotopy Theory*, (Academic Press, NY, 1956); N. Steenrod, *The topology of Fiber Bundles*, (Princeton Univ. Press, NJ 1951).
14. E. Witten, *Phys. Lett.* **B117**, (1982) 324.
15. S. Elitzur and V. P. Nair, *Nucl. Phys.* **B243** (1984) 205.
16. S. Okubo, H. Zhang, Y. Tosa, and R. E. Marshak, *Phys. Rev.* **D37**, (1988) 1655; H. Zhang, S. Okubo, and Y. Tosa, *Phys. Rev.* **D37**, (1988) 2946; H. Zhang and S. Okubo, *Phys. Rev.* **D38**, (1988) 1800; S. Okubo and H. Zhang, in *Perspectives on Particle Physics*, ed. S. Matsuda, T. Muta, and R. Sakaki, (World Scientific, 1989); A. T. Lundell and Y. Tosa, *J. Math. Phys.* **29**, (1988) 1795; S. Okubo and Y. Tosa, *Phys. Rev.***D40**, (1989) 1925; H. Zhang, *Z. Phys.* **C52**, (1991) 455.
17. See P. Goddard and D. Olive, *Rep. Prog. Phys.* **41**, (1978) 1375; P. Goddard, J. Nuyts, D. Olive, *Nucl. Phys.* **B125**, (1977) 1.
18. C. Dokos and T. Tomaras, *Phys. Rev.* **D21**, (1980) 2940.
19. G. Giacomelli, in *Monopoles in Quantum Field Theory*, (World Scientific, 1981).
20. H. Pagels, *Phys. Rev.* **D13** (1976) 343; W. Marciano and H. Pagels,*Phys. Rev.* **D14** (1976) 531.
21. E. Witten, *Phys. Lett.* **B86** (1979) 283; E. Tomboulis and G. Woo,*Nucl. Phys.* **B107** (1976) 221.
22. H. Zhang, *Phys. Rev.* **D36** (1987) 1868.
23. H. Zhang, *Preprint* **LBL-32531** June 1992.

Email: ZHANGHZ@SSCVX1.SSC.GOV